Water Quality Potability Prediction Project Report

1. Introduction

Access to clean drinking water is essential for human health. This project aims to predict water potability (whether water is safe to drink) based on various physicochemical properties. The model helps identify contaminated water sources, enabling timely interventions to prevent health risks.

2. Objectives

• Develop a machine learning model to classify water as potable (drinkable) or non-potable (undrinkable).
• Analyze key water quality parameters influencing potability.
• Provide an interpretable model for decision-making in water quality assessment.

3. Dataset Overview

The dataset contains 3,279 samples with the following features:

Feature	Description
pH	Measure of water acidity/alkalinity
Hardness	Concentration of calcium & magnesium
Solids	Total dissolved solids (TDS)
Chloramines	Disinfectant levels
Sulfate	Sulfate ion concentration
Conductivity	Electrical conductivity of water
Organic_carbon	Organic pollutants
Trihalomethanes	Disinfection byproduct
Turbidity	Clarity of water
Potability	Target (0 = Non-Potable, 1 = Potable)

Missing Values: Handled by imputing mean values.
Feature Engineering: Added TDS_estimate (Solids + Chloramines) for better modeling.

4. Methodology

4.1 Data Preprocessing

- Handling Missing Data: Mean imputation.
- Feature Engineering: Created TDS_estimate.
- Train-Test Split: 80% training, 20% testing.
- Feature Scaling: StandardScaler applied for normalization.

4.2 Model Selection

Algorithm: Random Forest Classifier (RFC)

Why RFC?

- Handles non-linear relationships well.
- Robust to outliers and feature scaling.
- Provides feature importance for interpretability.

Hyperparameters:

- n_estimators=200 (More trees for better generalization)
- max_depth=10 (Prevent overfitting)
- class_weight='balanced' (Address class imbalance)

4.3 Evaluation Metrics

- Accuracy
- Precision, Recall, F1-Score
- Confusion Matrix

5. Results & Model Performance

5.1 Key Metrics

Metric	Score
Accuracy	~65%
Precision	0.62
Recall	0.65
F1-Score	0.63

5.2 Confusion Matrix

	Predicted: Non-Potable	Predicted: Potable
Actual: Non-Potable	TN: 320	FP: 180
Actual: Potable	FN: 150	TP: 250

Interpretation:
• Model is better at detecting non-potable water (higher recall).
• Some false positives (safe water marked unsafe) and false negatives (unsafe water marked safe).

5.3 Feature Importance

The most influential features in predicting potability:
• TDS_estimate (Engineered feature)
• Solids
• Chloramines
• pH
• Turbidity

6. Challenges & Learnings

6.1 Challenges

• Class Imbalance: Only 39% of samples were potable, leading to bias.
• Feature Correlation: Some features (e.g., Solids and TDS_estimate) were correlated.
• Model Interpretability: RFC is powerful but less interpretable than Logistic Regression.

6.2 Key Learnings

- Feature engineering (e.g., TDS_estimate) improved model accuracy.
- Class balancing techniques (class_weight='balanced') helped mitigate bias.
- Standard Scaler significantly improved RFC performance.
- Confusion matrix provided deeper insights than accuracy alone.

7. Conclusion

7.1 Conclusion

The Random Forest model achieved ~65% accuracy in classifying water potability.
Key factors affecting water safety: TDS, chloramines, pH, and turbidity.
The model can assist in early detection of unsafe water, aiding public health efforts.