Hybrid Water Quality Assessment Using K-Means, Random Forest, SVM, and k-NN: A Case Study of the Mahakali River, Nepal

Tek Raj Bhatt
Student ID: 250069, CUID: 16544288
MSc. Data Science and Computational Intelligence
Softwarica College of IT and E-commerce (Coventry University)
250069@softwarica.edu.np

Abstract—**This study presents a hybrid machine learning approach to water quality assessment in Nepal's Mahakali River, combining unsupervised pattern discovery with supervised classification. Using bi-monthly measurements of ten water quality parameters collected across three districts over three years, the research employs K-means clustering to identify natural patterns, followed by Random Forest, SVM, and k-NN classification algorithms to predict quality levels. The framework aims to support evidence-based water resource management in the Mahakali River basin.**

I. Introduction

Water quality monitoring is critical for public health in Nepal's Mahakali River basin, serving communities across Baitadi, Dadeldhura, and Kanchanpur districts. This study applies machine learning to discover patterns in multi-dimensional water quality data and develop predictive classification models. Using real-world data collected by Oxfam under the TROSA programme (ten parameters, bi-monthly, three years, fifteen locations), the research bridges unsupervised pattern discovery with supervised classification for automated water quality assessment.

II. Objective

- Discover natural water quality patterns through PCA and K-means clustering.
- Generate quality labels using WHO/Nepal standards and Water Quality Index (WQI) methodology.
- Develop and compare supervised classification models (Random Forest, SVM, k-NN) to identify influential parameters for actionable water management insights.

III. Data Source

Water quality data from Oxfam's TROSA programme covering three Far-Western Nepal districts (Baitadi, Dadeldhura, Kanchanpur) with 15 monitoring stations (5 per district). Bi-monthly measurements over three years include: **Physical** (Temperature, pH, Hardness); **Chemical** (Chloride, Ammonia, Phosphate, Nitrate, Iron, Free Residual Chlorine); **Biological** (Coliform). No Personally Identifiable Information (PII); data provided with appropriate permissions for academic research.

IV. Methodology

Phase 1 - Unsupervised Discovery: Handle missing values, Z-score normalization, EDA, PCA (85-90% variance), K-means clustering (elbow method and silhouette analysis for optimal cluster selection), cluster profiling with radar plots and statistical summaries.

Phase 2 - Label Generation: Standards-based labeling (WHO/Nepal criteria) or WQI calculation ($WQI = \sum W_i \times Q_i$) to categorize samples as Excellent/Good/Poor/Very Poor quality.

Phase 3 - Supervised Classification: Feature engineering (interaction features, seasonal indicators), train and compare three classifiers (Random Forest, SVM with RBF kernel, k-NN), handle class imbalance using SMOTE if needed.

Phase 4 - Evaluation: 5-fold stratified cross-validation, performance metrics (Accuracy, Precision, Recall, F1-score, AUC-ROC), confusion matrices for comprehensive model comparison.

Phase 5 - Insights: Feature importance analysis using Random Forest and SHAP values, spatial-temporal quality mapping, recommendations for monitoring station optimization.

V. Expected Outcomes

- Identification of distinct water quality clusters with clear characterization
- Classification models achieving more than 80% accuracy with comprehensive performance metrics
- Feature importance ranking (expected: coliform, pH, ammonia, nitrate as top predictors)
- Spatial-temporal quality maps showing district-level and seasonal patterns
- Actionable recommendations for monitoring station optimization and water resource management

VI. Significance

This research demonstrates a hybrid ML framework for environmental monitoring, providing actionable insights for water resource management authorities and contributing to Oxfam's TROSA through evidence-based research.