Intermediate Report and Code Analysis

Ravi Kant Gupta, Ayushi Choyal, Shouryavi Awasthi

October 17, 2025

1 Project Status and Milestone Overview

Milestone: Intermediate Report & Code (Due: 17 Oct)
Goal: Implement advanced models, integrate spatio-temporal data, and establish analytical conclusions for both project objectives (Classification and Regression).

This phase focused on advancing the baseline models and successfully overcoming the technical challenge of integrating the complex UCI spatio-temporal dataset (`water_dataset.mat`).

2 Classification: Advanced Model Results (Water Potability)

The advanced models (Random Forest and XGBoost), configured with `class_weight='balanced'` to address the $61\% : 39\%$ imbalance, were deployed against the Logistic Regression Benchmark (F1-Score $= 0.4666$).

2.1 Model Performance Summary

Table 1: Advanced Classification Model Performance

Model	**F1-Score** (1)	**ROC-AUC**	**Improvement vs. Baseline (0.4666)**
Logistic Regression (Baseline)	0.4666	0.5475	—
Random Forest	0.4495	**0.6765**	-3.7% F1
XGBoost	0.4551	0.6178	-2.5% F1

2.2 Analytical Conclusions on Classification

1. **Underperformance Confirmation:** Neither the Random Forest nor the XGBoost model achieved an F1-Score greater than the simple linear baseline. This confirms the initial hypothesis that the decision boundary is complex and non-linear.
2. **High Potential / Threshold Problem (Major Insight):** Despite the low F1-Score, the Random Forest model achieved a high **ROC-AUC of 0.6765**. This is a critical finding: the model has a strong ability to rank positive and negative cases but is failing due to the standard classification threshold (default $= 0.5$).
3. **Feature Importance:** Feature analysis confirmed the primary drivers, which will guide further modeling: `ph` (0.1507), `Sulfate` (0.1497), and `Hardness` (0.1187).

3 Regression: Spatio-Temporal Baseline (pH Forecasting)

This phase successfully overcame the file access hurdle and established a benchmark for predicting pH over time using the multi-dimensional UCI dataset.

3.1 Data Integration and Baseline Model

- **Data Restructuring:** The multi-dimensional Location × Feature × Time data structure was successfully flattened into a standard tabular format (15,651 observations) suitable for supervised learning.

- **Model Implemented:** Simple **Linear Regression** was run on the scaled data to establish the benchmark.

3.2 Benchmark Performance

Table 2: Linear Regression Baseline Metrics (pH Forecasting)

Metric	Score	Interpretation
$\mathbf{R^2}$ Score	**0.8329**	The linear model explains over 83% of the variance in the target pH value. This is an extremely strong result for a baseline model.
RMSE	**0.0120**	The average prediction error is very low, confirming the predictive features are highly correlated with pH.

4 Final Conclusion and Future Work

Primary Conclusion: The Intermediate phase confirmed that while highly predictive features exist in the spatio-temporal data (Regression $\mathbf{R^2} = \mathbf{0.8329}$), the core potability task (Classification) is bottlenecked by the difficulty of separating overlapping classes.

4.1 Final Phase Action Plan (Due: 3 Nov)

i. **Classification Fix:** Implement **Classification Threshold Tuning** (e.g., G-Mean method) on the Random Forest model to convert the high ROC-AUC (0.6765) into an F1-Score definitively greater than the 0.4666 benchmark.

ii. **Regression Advancement:** Implement a specialized non-linear model, such as a **Simple RNN/LSTM** (Recurrent Neural Network), to attempt to capture the remaining 17% of variance in the pH forecast.

iii. **Report Finalization:** Compile all results, reflections, and declarations into the final `REPORT.pdf` and `slides/`.