Baseline Model Implementation Report: Water Potability Classification

Project Context

Milestone: Baseline Model Implementation (Due: October 6)
Dataset: Water Potability Classification (water_potability.csv)
Model: Logistic Regression (with Class Weighting)
Goal: Establish a performance benchmark for subsequent advanced modeling efforts.

1. Methodology and Data Preparation

The objective of the baseline is to provide a quick, interpretable, and reproducible starting point. The entire process used the cleaned dataset prepared during the EDA phase.

1.1 Data Preprocessing Summary

1. **Missing Value Imputation:** The columns ph, Sulfate, and Trihalomethanes were imputed using the **median** value of each respective column to maintain data integrity while avoiding skewness from mean imputation.
2. **Feature and Target Separation:** The data was separated into the feature matrix () and the binary target vector (, 'Potability').
3. **Stratified Splitting:** Due to the inherent **class imbalance** (61% Not Potable vs. 39% Potable), a stratified 80/20 train-test split was performed (, , ,) to ensure both sets maintained the original class ratio.
4. **Feature Scaling: StandardScaler** was fitted on the training data and then applied to both the training and test sets. This standardization process ensures that features with larger numerical ranges (like Solids) do not unfairly dominate the distance calculations in the Logistic Regression algorithm.

1.2 Model Selection and Imbalance Mitigation

Logistic Regression was chosen as the baseline model due to its simplicity and interpretability. Crucially, the model was configured to address the class imbalance:

- **Model:** LogisticRegression(class_weight='balanced', ...)
- **Imbalance Handling:** The **class_weight='balanced'** parameter was activated. This setting automatically adjusts weights inversely proportional to the class frequencies, forcing the model to heavily penalize misclassifications of the minority class (Potable water). This step was essential, as the standard (unbalanced) Logistic Regression model previously failed by predicting the majority class exclusively.

2. Baseline Model Performance Analysis

The balanced Logistic Regression model was evaluated on the unseen test set (20% of the data, 656 samples).

2.1 Key Evaluation Metrics

The results show a clear trade-off between the model's ability to correctly identify the positive class (Recall) and its reliability when making that positive prediction (Precision).

Metric	Score	Interpretation
Accuracy	0.53	The model is correct 53% of the time overall. This is only slightly better than random guessing (0.50).
Precision (Potable=1)	0.42	When the model predicts water is **Potable**, it is correct **42%** of the time.
Recall (Potable=1)	0.53	The model correctly identifies **53%** of all truly Potable samples.
F1-Score (Potable=1)	**0.4666**	The harmonic mean of Precision and Recall for the minority class. This serves as the established **quantitative benchmark**.
ROC-AUC Score	**0.5475**	The model's ability to distinguish between the two classes. A score close to 0.50 indicates very weak separability.

2.2 Detailed Classification Breakdown

Class	Precision	Recall	F1-Score	Support
0 (Not Potable)	0.64	0.52	0.57	400
1 (Potable)	0.42	0.53	0.47	256
Weighted Avg.	0.55	0.53	0.53	656

Analysis of Bias vs. Predictive Power:

1. **Trade-off Confirmation:** The model sacrificed high accuracy on the majority class (Precision 0.64, Recall 0.52) to achieve decent Recall on the minority class (Recall 0.53). This confirms that the class_weight='balanced' parameter successfully forced the model away from its initial, useless accuracy bias.

2. **Linear Limitation:** The poor F1-Score (0.47) and low ROC-AUC (0.5475) indicate that **potability is not linearly dependent** on the set of water quality parameters. The simple linear boundary created by Logistic Regression is insufficient to separate the two highly overlapping classes, validating the hypothesis from the EDA phase.

3. Conclusion and Future Work

3.1 Establishing the Benchmark

The primary deliverable of this milestone—the **Baseline Model Benchmark**—is established at an **F1-Score of 0.4666** for the Potable class. This score quantifies the current predictive challenge and sets the minimum performance target for all subsequent, more complex models.

3.2 Next Steps (Intermediate Report: 17 Oct)

The results strongly suggest a need to move to models capable of learning complex, non-linear feature interactions:

1. **Advanced Classification Models:** The next phase will focus on non-linear ensemble methods such as the **Random Forest Classifier** and **Gradient Boosting (e.g., XGBoost)**. These models inherently handle non-linear data better and will be implemented with class_weight='balanced' to maintain the learned balance.

2. **Spatio-Temporal Data Integration:** The Intermediate phase will also begin the challenging task of loading and structuring the UCI **.mat** file dataset for time-series forecasting of pH, requiring libraries like scipy.io and potentially specialized time-series modeling techniques (e.g., recurrent or state-space models) [cite: uploaded:water+quality+prediction-1.zip/README.docx].

3. **Feature Importance:** Subsequent models will be analyzed for feature importance (e.g., using SHAP or feature coefficients) to gain physical interpretability, a crucial objective of the proposal [cite: uploaded:Water_Quality_Proposal_Kurukshetra (1).docx].