HAB (Harmful Algal Bloom) Prediction Models

This repository contains machine learning models and an interactive dashboard for predicting Harmful Algal Blooms (HAB) in water systems. The implementation includes Random Forest and XGBoost models trained on 9 key predictors, along with a user-friendly dashboard for real-time predictions and sensitivity analysis.

Repository Structure

```
HAB/
│
├── Data_2/                    # Training data directory
│   └── UCDavis_Data_for_Analysis2_220days_ToxinProdMicrocystis.csv
│
├── Model/                     # Saved model pipelines
│   ├── rf_pipeline.joblib
│   └── xgb_pipeline.joblib
│
├── environment.yml           # Environment configuration
├── Testing_RF_XGB_for_reproducability_V3.py    # Model training script
└── HAB_Dashboard_9features_RFandXGB.py         # Interactive dashboard
```

Features

The models use the following 9 predictors, as recommended by Gourab Saha:

- Q: Flow Rate (m^3/s)
- Q_1m: Flow Rate 1 Month Ago (m^3/s)
- WaterTemp: Water Temperature ($^\circ C$)
- DO: Dissolved Oxygen (mg/L)
- pH: pH Level
- SpecCond.: Specific Conductance ($\mu S/cm$)
- OrgC: Organic Carbon (mg/L)
- OrgN: Organic Nitrogen (mg/L)
- PO4: Phosphate (mg/L)

Model Specifications

Random Forest

```
RandomForestClassifier(
    n_estimators=200,
    max_depth=20,
    min_samples_split=10,
    min_samples_leaf=4,
    max_features='log2',
    bootstrap=False,
    class_weight='balanced',
    criterion='log_loss',
    max_samples=None,
    random_state=42
)
```

XGBoost

```
XGBClassifier(
    learning_rate=0.01,
    n_estimators=100,
    max_depth=9,
    min_child_weight=1,
    gamma=0,
    subsample=0.8,
    colsample_bytree=0.6,
    alpha=0.5,
    lambda_=0.5,
    scale_pos_weight=1,
    base_score=0.5,
    booster='gbtree',
    eval_metric='logloss',
    random_state=42
)
```

Installation

1. Create and activate the conda environment:
conda env create -f environment.yml
conda activate HabsEnv_2025

Running the Dashboard

The dashboard provides an interactive interface for HAB prediction and sensitivity analysis. You don't need to retrain the models to use the dashboard as it uses pre-trained model pipelines.

The dashboard will be available at `http://localhost:5006` in your web browser.

Training the Models (Optional)
If you want to retrain the models:

This script will:
- Load and preprocess the data
- Train Random Forest and XGBoost models
- Generate performance metrics
- Save the model pipelines

Dashboard Features

The interactive dashboard includes:
- Real-time prediction using both Random Forest and XGBoost models
- Probability estimates for HAB occurrence

- Model comparison visualization
- Sensitivity analysis for each input parameter
- Interactive sliders for parameter adjustment