Leveraging Machine Learning for Sustainable Water Quality Modeling

WQI Classification • Spatio-Temporal Analysis

Aakash Vardhan Madabushi (018291663)
Kruthi Shamanna (018320770)
Sai Sushma Maddali (018195775)
Vedika Sumbli (018305937)

Agenda

- Introduction
- Dataset Overview
- Data Preprocessing
- WQI Computation
- Feature Engineering
- Machine Learning Pipeline
- Experimental Setup
- Results & Model Comparison
- Streamlit Web Application
- Geospatial Insights
- Challenges & Lessons Learned
- Future Work
- Conclusion
- Appendix (Rubric Evidence, Tools, CRediT)

Problem Statement

Water quality monitoring is essential for public health and environmental safety.

Manual assessment is slow, costly, and inconsistent between stations.

California has ~30,000 sampling stations and large heterogeneous data.

Need an automated ML-based system to classify: **Good — Moderate — Poor— Very Poor WQI.**

Motivation

California water datasets are massive and multi-dimensional.

Agencies need early detection of pollution trends.

Manual WQI calculation is time-consuming.

Machine Learning can:

- Automate WQI prediction
- Discover hidden patterns
- Provide actionable insights via dashboards
- Align with **UN SDG 6: Clean Water and Sanitation**

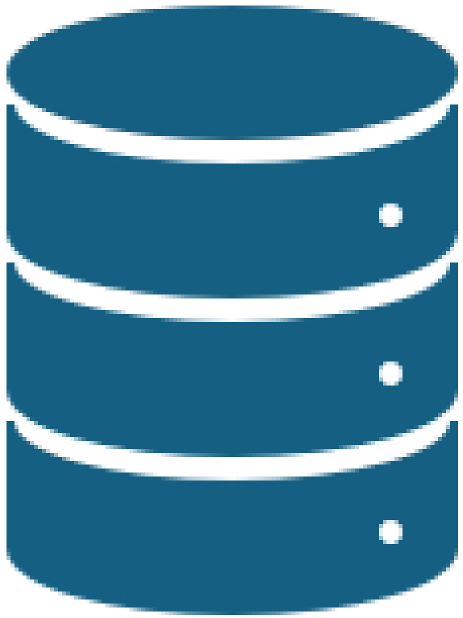
Dataset Overview – California DWR Field Results

Source: California Department of Water Resources (DWR)

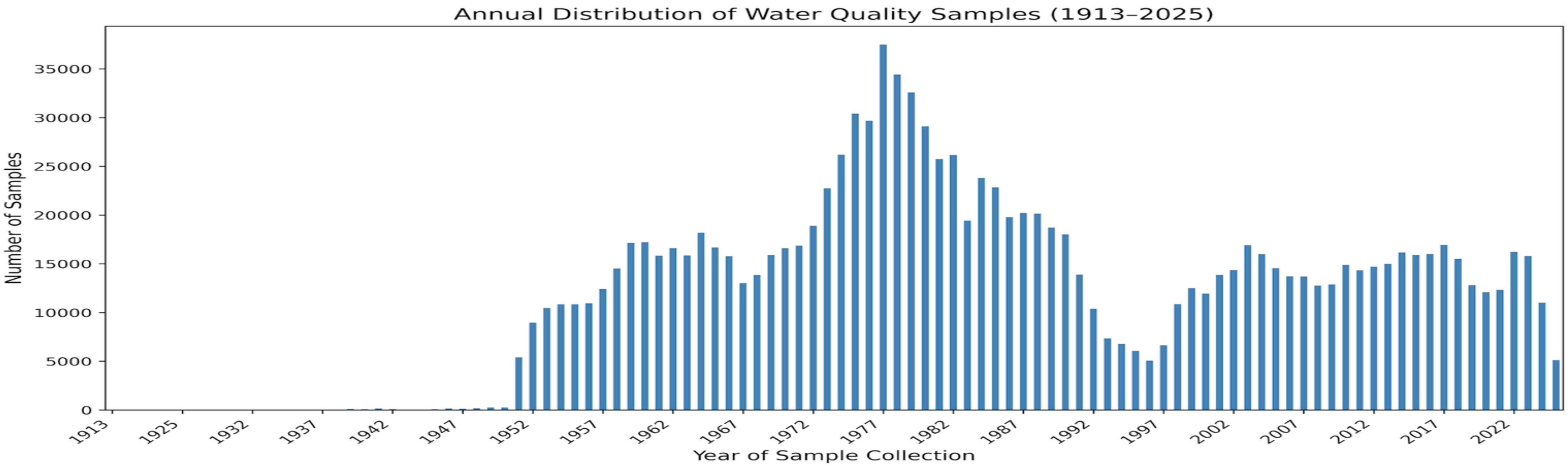
Access: https://data.ca.gov/

Type: Field water-quality measurements (surface and groundwater)

Field Results dataset from California DWR provides in-situ water quality measurements (e.g., temperature, turbidity, dissolved oxygen, conductivity) collected at monitoring stations for statewide assessment.



Dataset Characteristics

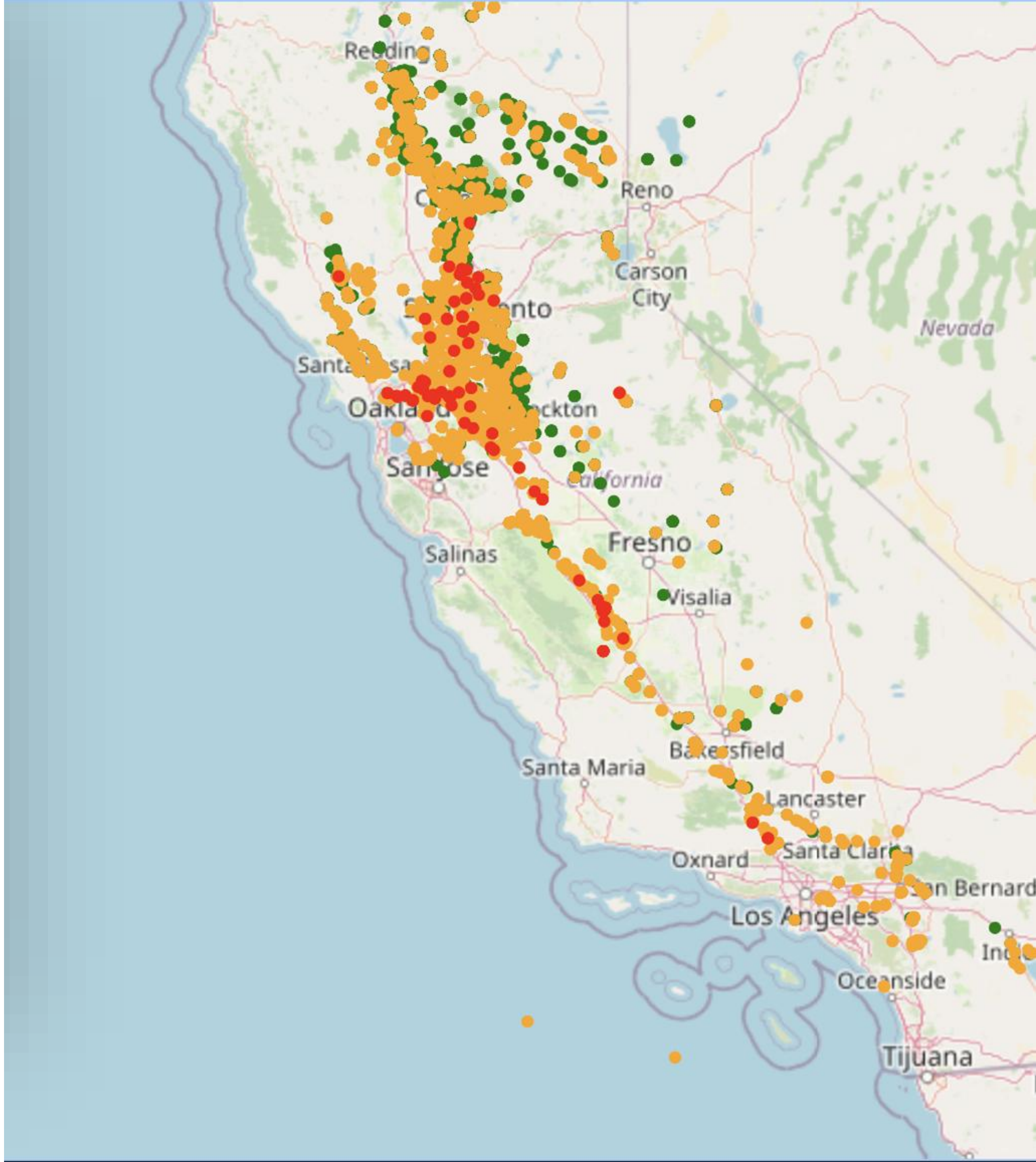


- ❖ **Time Range (Original):** 1913 to 2025
- ❖ **Dimensions (Original)**: 1.2 million+ records , 22 features
- ❖ **No. of Water Monitoring Stations**: 30,000 approx.
- ❖ **No. of Counties**: 58
- ❖ **No. of parameters measured**: 65 (physicochemical and environmental)
- ❖ **Time Range for Modeling** : 2000 to 2025
- ❖ **Dimensions (Modeling)**: ~50000, 12 features

Snapshot of Data

Field	Value
station_id	12
station_name	H.O. Banks Headworks
station_number	KA000331
full_station_name	Delta P.P. Headworks at H.O. Banks PP
station_type	Surface Water
latitude	37.8019
longitude	-121.6203
status	Public, Review Status Unknown
county_name	Alameda
sample_code	OM0168A0001
sample_date	1/4/1968 7:45
sample_depth	1
sample_depth_units	Feet
parameter	Dissolved Oxygen
fdr_result	9.2
fdr_reporting_limit	0.2
uns_name	mg/L
mth_name	EPA 360.2 (Field)

The dataset provides comprehensive, long-term environmental monitoring data that serves as a foundation for sustainable water-quality prediction using machine learning



Data Preparation Steps

Data Loading and Inspection

- *Inspect shape, columns distributions*
- *Missing values*
- *Visualize sample trends*

Filtering and Cleaning

- Filter by year and county, drop sparse and constant columns,
- Select core parameters

Standardization and Unit Conversion

- *Normalize depth and unit formats,*
- *Convert values to standard units,*
- *Create composite parameter-unit ID*

Missing Value Imputation & Outlier Handling

- *Impute coordinates and limits,*
- *Cap and fill depth values,*
- *Detect and clean outliers*

Water Quality Index (WQI) Calculation

- *Compute sub-indices for 5 parameters,*
- *Apply scaling and piecewise mapping,*
- *Aggregate into WQI via weighted average,*
- *Classify water quality: Good, Moderate, Poor, Very Poor*

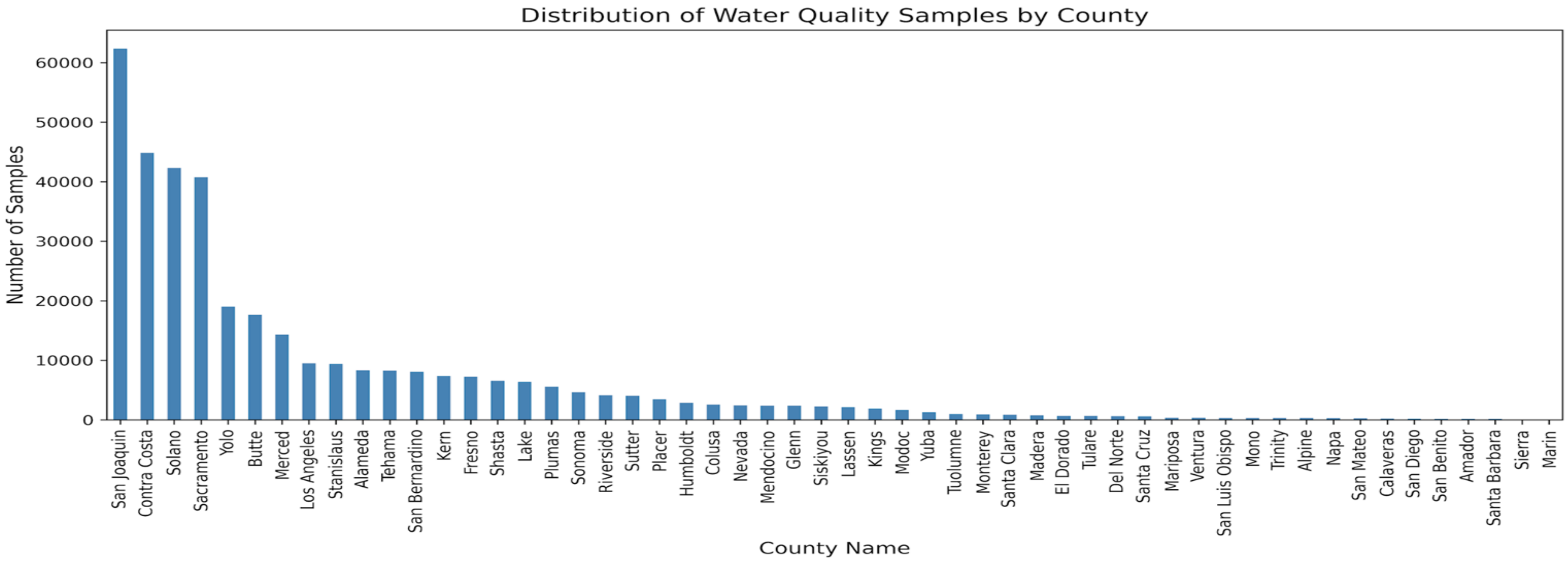
Pivoting & Dataset Consolidation

- *Reshape data to wide format,*
- *One row per sample; parameters as columns,*
- *Preserve station metadata for analysis*

Exploratory Data Analysis

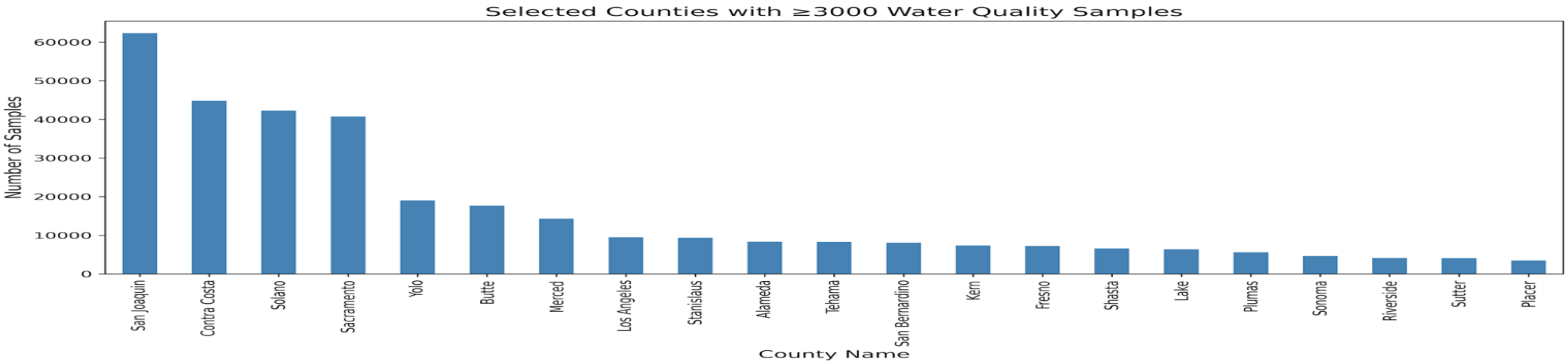
- Exploring data distribution
- Parameter Analysis
- Missing values Analysis
- Outlier Analysis
- Spatio-Temporal Analysis

Data Distribution

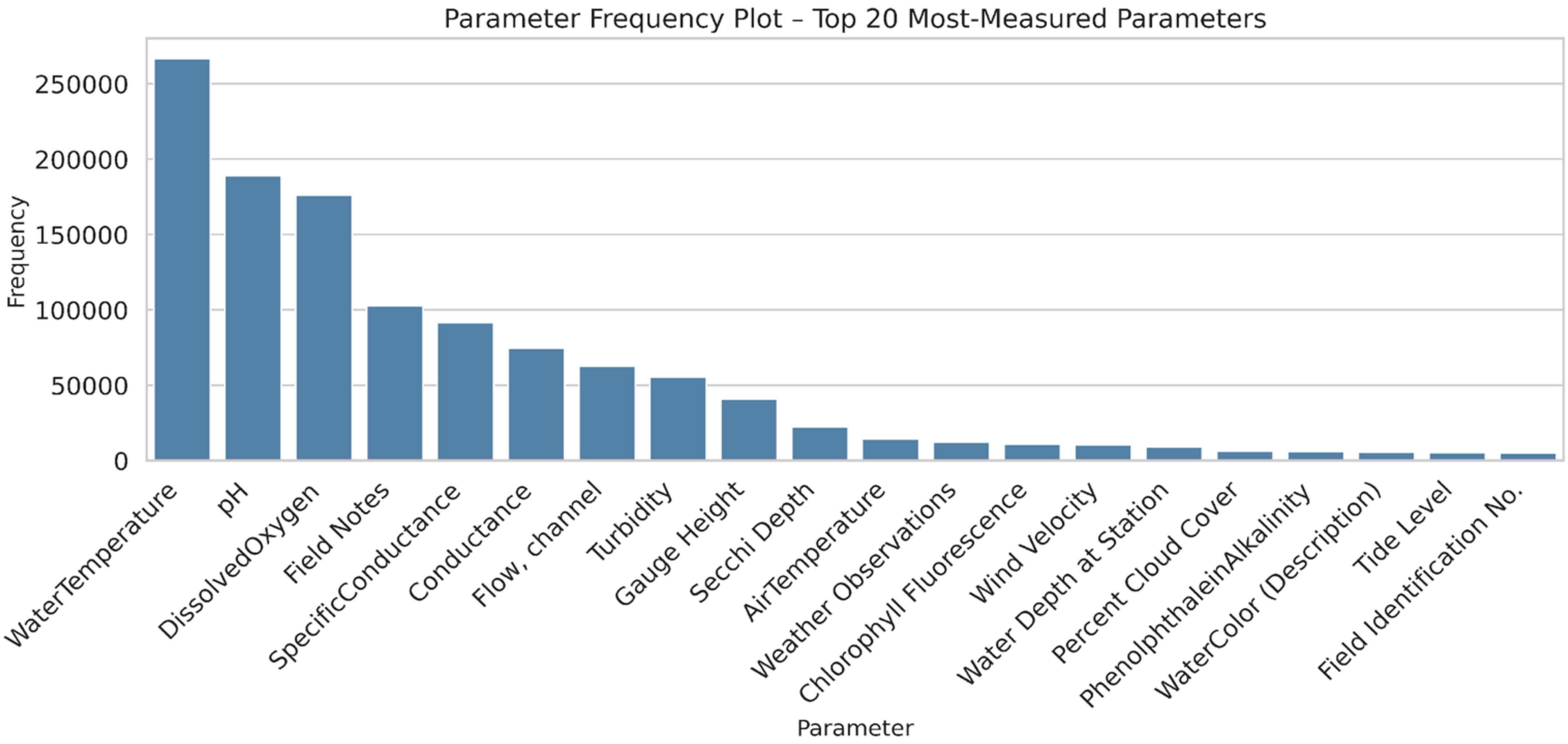


Original Distribution

After Filtering



Frequency distribution of leading physicochemical parameters



Handling Missing Values

✓ **Structured Imputation Strategy**

- *Designed to retain maximum samples for model training.*
- *Different methods applied depending on data type.*

✓ **Parameter Selection (Coverage-Based)**

- *Parameters with >80% missing values removed.*
- *Five core parameters retained (70–98% coverage):*
- *pH, Dissolved Oxygen, Turbidity, Specific Conductance, Water Temperature.*

✓ **Geographic Location Imputation**

- *Missing station coordinates estimated using county-level averages.*
- *Preserves regional context instead of using statewide averages.*

✓ **Water Quality Parameter Imputation**

- *Remaining gaps filled with median values.*
- *Median chosen to reduce influence of outliers.*
- *Ensures realistic distributions before WQI calculation and model training.*

Unit Standardization and Cleaning

Before Standardization

Parameter	Units Found
(Bottom) DissolvedOxygen	mg/L, % Saturation
(Bottom) SpecificConductance	uS/cm@25 °C, µmhos/cm@25°C
(Bottom) WaterTemperature	°C
(Bottom) Chlorophyll Fluorescence	ug/L of Chl, RFU
(Bottom) Turbidity	N.T.U., F.N.U.
(Bottom) pH	pH Units
Carbon Dioxide	mg/L
Chlorophyll Fluorescence	RFU, ug/L of Chl
Chlorophyll Volume	mL
Discharge	cfs
DissolvedOxygen	mg/L, % Saturation, %
ElectricalConductance	uS/cm, uS/cm@25 °C
Flow, channel	cfs, Gallons
Redox Potential	mV
Secchi Depth	Meters, Feet, Centimeters
SoilRedox Potential	mV
SpecificConductance	uS/cm@25 °C, µmhos/cm@25°C
SpecificConductance (EC w/time)	uS/cm@25 °C
Turbidity	N.T.U., F.N.U.
Turbidity (w/time)	N.T.U.
WaterTemperature	°C, , °F
WaterTemperature (w/time)	°C
pH	pH Units
pH (w/time)	pH Units

After Standardization

Parameter	Units Found
(Bottom) DissolvedOxygen	mg/L
(Bottom) SpecificConductance	µS/cm
(Bottom) WaterTemperature	°C
(Bottom) Chlorophyll Fluorescence	µg/L
(Bottom) Turbidity	NTU
(Bottom) pH	pH units
Carbon Dioxide	mg/L
Chlorophyll Fluorescence	µg/L
Chlorophyll Volume	µg/L
Discharge	m^3/s
DissolvedOxygen	mg/L
ElectricalConductance	µS/cm
Flow, channel	m^3/s
Redox Potential	mV
Secchi Depth	m
SoilRedox Potential	mV
SpecificConductance	µS/cm
SpecificConductance (EC w/time)	µS/cm
Turbidity	NTU
Turbidity (w/time)	NTU
WaterTemperature	°C
WaterTemperature (w/time)	°C
pH	pH units
pH (w/time)	pH units

Outlier Detection and Handling

- ❖ Implausible readings (e.g., negative temperature, pH > 14, DO > 20 mg/L) were flagged and replaced with missing values.
- ❖ 296 records (0.1%) removed to preserve environmental realism and prevent model bias.
- ❖ Valid parameter ranges defined using EPA, WHO, and APHA guidelines.

Examples:

- pH range: 6.5–8.5
- Dissolved Oxygen: ≤ 20 mg/L
- Ensured scientific validity for WQI computation and ML model training.

Parameter	Valid Range	Reference Source
Dissolved Oxygen (mg/L)	0–20	WHO [19], EPA [18]
Water Temperature (°C)	-2–50	EPA [18], WHO [19]
pH (pH units)	0–14 (optimal 6.5–8.5)	EPA [17], WHO [19]
Specific Conductance (µS/cm)	0–50,000	USGS [20]
Turbidity (NTU)	0–1000	WHO [19]
Secchi Depth (m)	0–50	OECD [21]
Chlorophyll Fluorescence (µg/L)	0–1000	UNESCO [22]
Redox Potential (mV)	-500–1000	APHA [23]
Carbon Dioxide (mg/L)	0–200	Wetzel [24]

Table showing Valid Environmental Ranges for Core Water-Quality Parameters

Pivoting Data to Wide Format

Reshape multiple measurements per sampling event into a single row for analysis.

Before Pivoting

Field	Value
station_id	12
station_name	H.O. Banks Headworks
station_number	KA000331
full_station_name	Delta P.P. Headworks at H.O. Banks PP
station_type	Surface Water
latitude	37.8019
longitude	-121.6203
status	Public, Review Status Unknown
county_name	Alameda
sample_code	OM0168A0001
sample_date	1/4/1968 7:45
sample_depth	1
sample_depth_units	Feet
parameter	Dissolved Oxygen
fdr_result	9.2
fdr_reporting_limit	0.2
uns_name	mg/L
mth_name	EPA 360.2 (Field)

After Pivoting

Field	Value
station_id	1
station_name	AMERICAN
station_number	A0714010
full_station_name	American River at Water Treatment Plant
station_type	Surface Water
latitude	38.5596
longitude	-121.4169
county_name	Sacramento
sample_code	C0114B0005
sample_date	1/6/2014 12:14
year	2014
sample_depth_meter	1
DissolvedOxygen_mg/L	12.18
SpecificConductance_µS/cm	66
Turbidity_NTU	2.28
WaterTemperature_°C	10.18
pH_pH units	7.6

Preparing the target: WQI Class

Compute Sub-Indices (qi)

$$
\begin{aligned}
q_{\text{DO}} &= \min\left(100, \frac{\text{DO}}{14} \times 100\right) \\
q_{\text{pH}} &= \begin{cases} 100 & \text{if } 6.5 \leq \text{pH} \leq 8.5 \\ 100 - 10|\text{pH} - 7.5| & \text{otherwise} \end{cases} \\
q_{\text{Cond}} &= \max\left(0, 100 - \frac{\text{Cond}}{1500} \times 100\right) \\
q_{\text{Turb}} &= \max\left(0, 100 - \frac{\text{Turb}}{100} \times 100\right) \\
q_{\text{Temp}} &= \max\left(0, 100 - |\text{Temp} - 20| \times 5\right)|
\end{aligned}
$$

Apply Weights

Weights (***Delphi method***):
DO 0.3, **pH** 0.2,
Conductance 0.2,
Turbidity 0.2, **Temperature 0.1**

Calculate WQI

$$\text{WQI} = \frac{\sum w_i q_i}{\sum w_i}$$

Classify Water Quality

Good: $\text{WQI} \geq 80$
Moderate: $50 \leq \text{WQI} < 80$
Poor: $25 \leq \text{WQI} < 50$
Very Poor: $\text{WQI} < 25$

Feature Engineering

Domain-specific features created to improve prediction accuracy and model interpretability.

Temporal Seasonality (Cyclic Encoding)

- Months encoded with sine/cosine functions to capture yearly cycles.
- Ensures January and December are treated as adjacent, reflecting true seasonal continuity.

$$\mathrm{Month}_{\sin} = \sin\left(\frac{2\pi m}{12}\right)$$

$$\mathrm{Month}_{\cos} = \cos\left(\frac{2\pi m}{12}\right)$$

Physical Interaction Feature (DO–Temp Ratio)

- Captures inverse relationship between water temperature and dissolved oxygen.
- High ratio → healthy water; low ratio → possible pollution or biological stress.

$$\mathrm{DO\_Temp\_Ratio} = \frac{\mathrm{DO}}{\mathrm{Temp} + 1}$$

Modeling

Feature & Target Selection
- Input features: numeric + categorical.
- Target: WQI_Class (encoded as Good=0, Moderate=1, Poor=2)

Train–Test Split
- 80/20 split with stratification to handle class imbalance.
- Split is done based on “station_id”

Preprocessing Pipelines
- Numeric: Standard scaling
- Categorical: One-Hot Encoding
- Target: Label Encoding

Handling Class Imbalance
- SMOTE Oversampling
- Applied after preprocessing to balance minority classes.

Model Training & Comparison
Classifiers tested:
- XGBoost
- Random Forest
- Logistic Regression
- SVM (RBF Kernel)

Evaluation
- Accuracy, Precision, Recall, F1 (macro), Matthews Correlation Coefficient, ROC-AUC, PR curves , Cohen’s Kappa

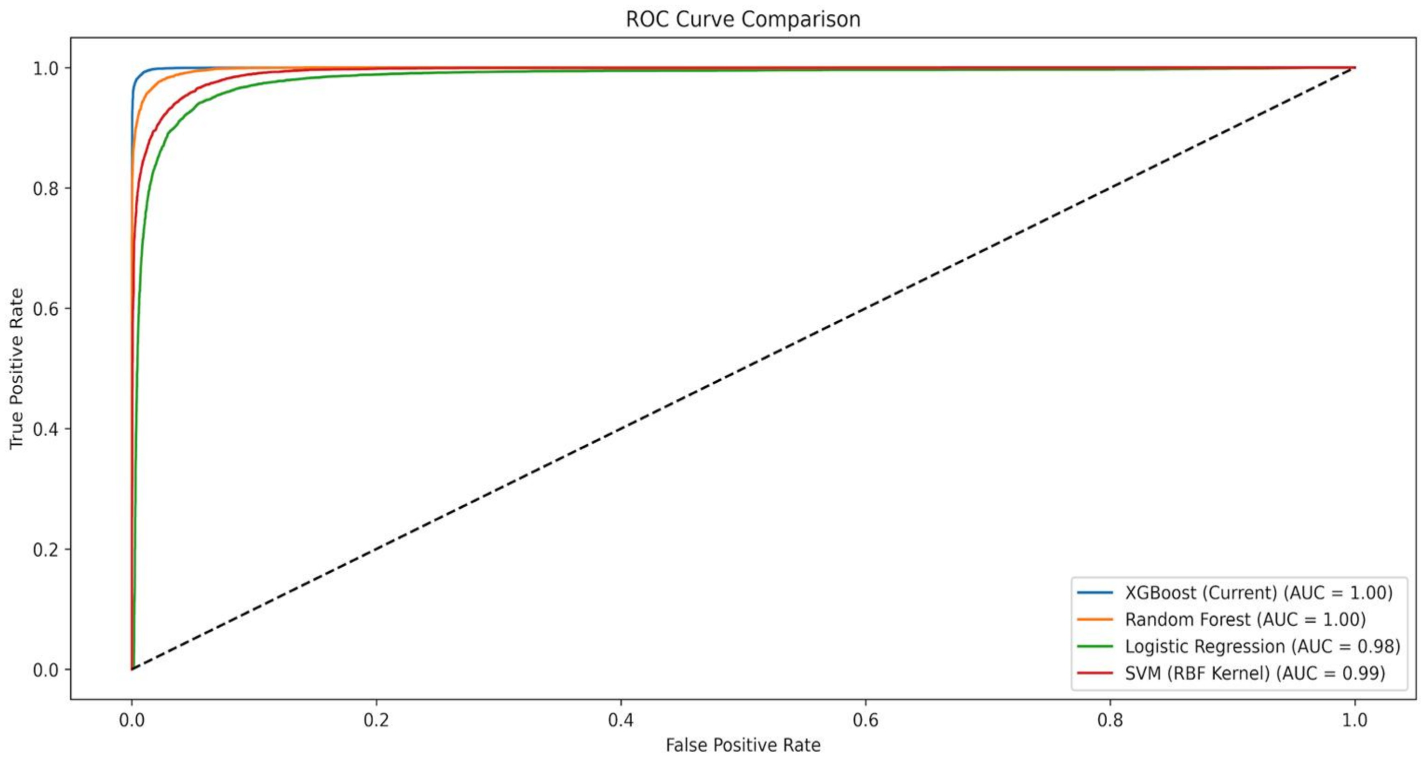
Category	Features	Description
Core Physicochemical	DissolvedOxygen_mg/L, pH_pH units, Turbidity_NTU, SpecificConductance_µS/cm, WaterTemperature_°C	Fundamental water quality indicators measuring chemistry and biology.
Station / Measurement	sample_depth_meter, station_type	Metadata about how/where the sample was collected.
Engineered Feature	DO_Temp_Ratio	Captures inverse DO–Temperature relationship (oxygen saturation proxy).
Temporal (Cyclic)	Month_sin, Month_cos	Encodes seasonality to reflect yearly cycles in water chemistry.
Spatial	latitude, longitude	Geographic coordinates of monitoring stations.
Target Variable	WQI_Class	Water quality classification (Good, Moderate, Poor, Very Poor).

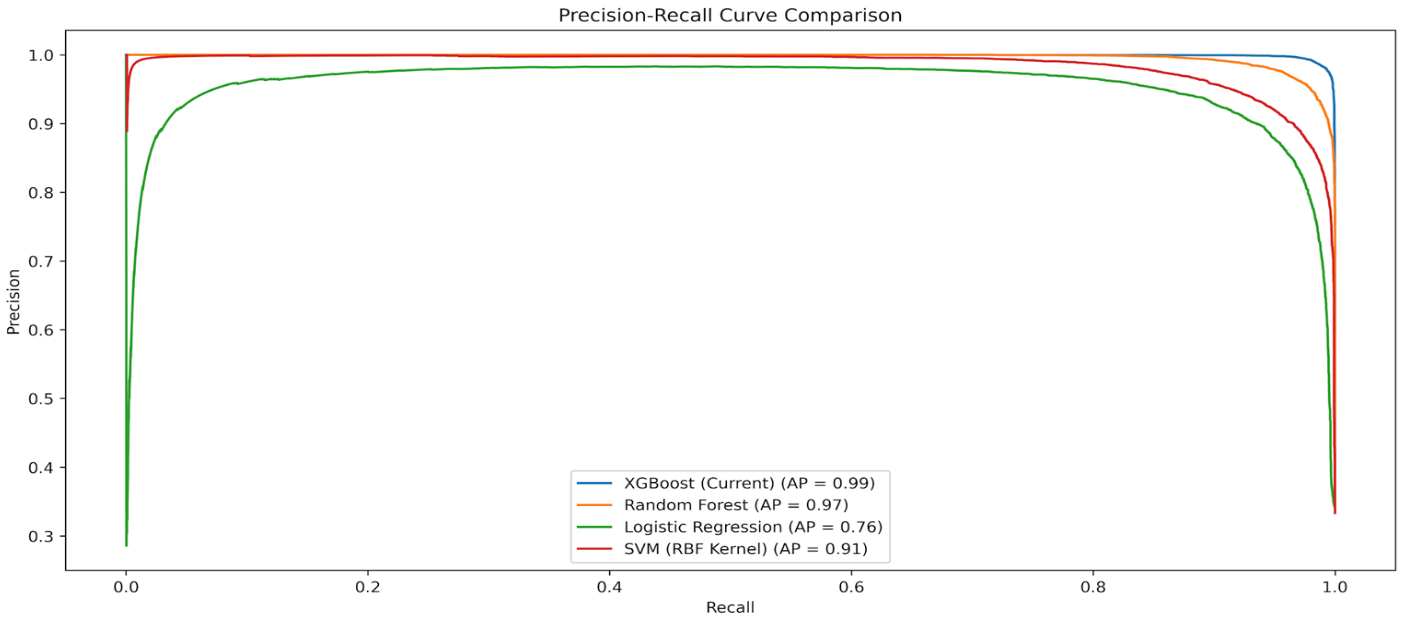
Features used for Training

Evaluation Metrics

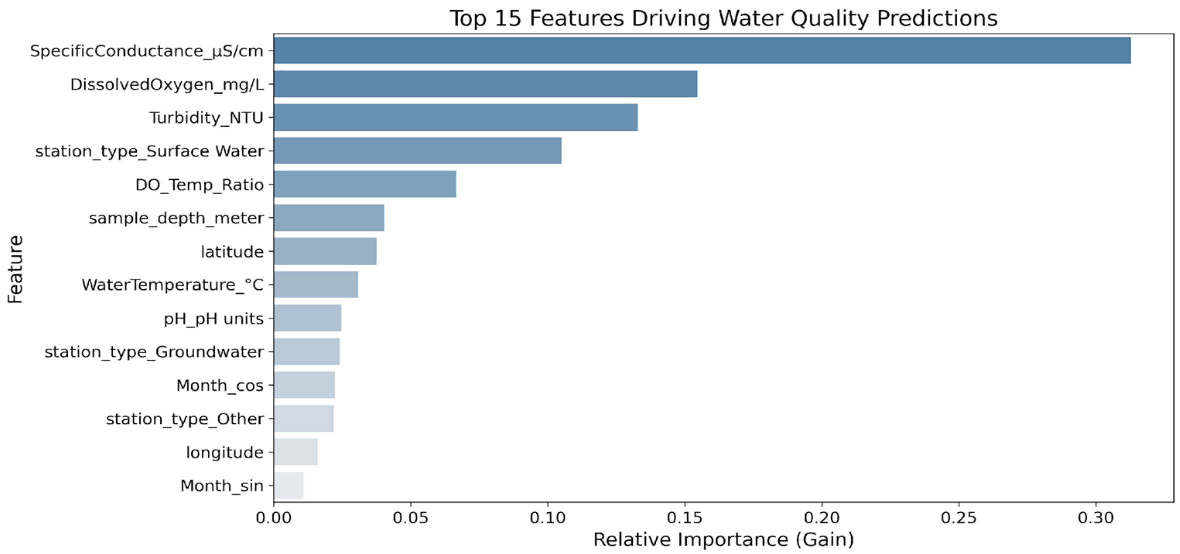
Model	Split	Class	Acc.	Prec.	Rec.	F1	Support
XGBoost	Train	Good	1.00	1.00	1.00	1.00	18,713
		Moderate	1.00	1.00	1.00	1.00	24,045
		Poor	1.00	1.00	1.00	1.00	199
	Test	Good	0.99	0.98	0.99	0.99	4,678
		Moderate	0.99	0.99	0.99	0.99	6,012
		Poor	0.99	0.88	0.90	0.89	50
Random Forest	Train	Good	0.98	0.98	0.99	0.98	18,713
		Moderate	0.98	0.99	0.98	0.99	24,045
		Poor	0.98	0.90	1.00	0.95	199
	Test	Good	0.97	0.96	0.97	0.96	4,678
		Moderate	0.97	0.98	0.96	0.97	6,012
		Poor	0.97	0.73	0.86	0.79	50
Logistic Reg.	Train	Good	0.92	0.91	0.96	0.93	18,713
		Moderate	0.92	0.96	0.89	0.92	24,045
		Poor	0.92	0.16	0.97	0.28	199
	Test	Good	0.92	0.91	0.95	0.93	4,678
		Moderate	0.92	0.96	0.89	0.92	6,012
		Poor	0.92	0.17	0.98	0.29	50
SVM (RBF)	Train	Good	0.93	0.90	0.96	0.93	18,713
		Moderate	0.93	0.97	0.90	0.93	24,045
		Poor	0.93	0.40	1.00	0.57	199
	Test	Good	0.93	0.90	0.96	0.93	4,678
		Moderate	0.93	0.97	0.90	0.93	6,012
		Poor	0.93	0.38	0.92	0.53	50

Model	Split	MCC	Kappa
XGBoost (Current)	Train	1.00	1.00
	Test	0.97	0.97
Random Forest	Train	0.97	0.97
	Test	0.93	0.93
Logistic Regression	Train	0.84	0.84
	Test	0.84	0.84
SVM (RBF Kernel)	Train	0.86	0.86
	Test	0.86	0.86

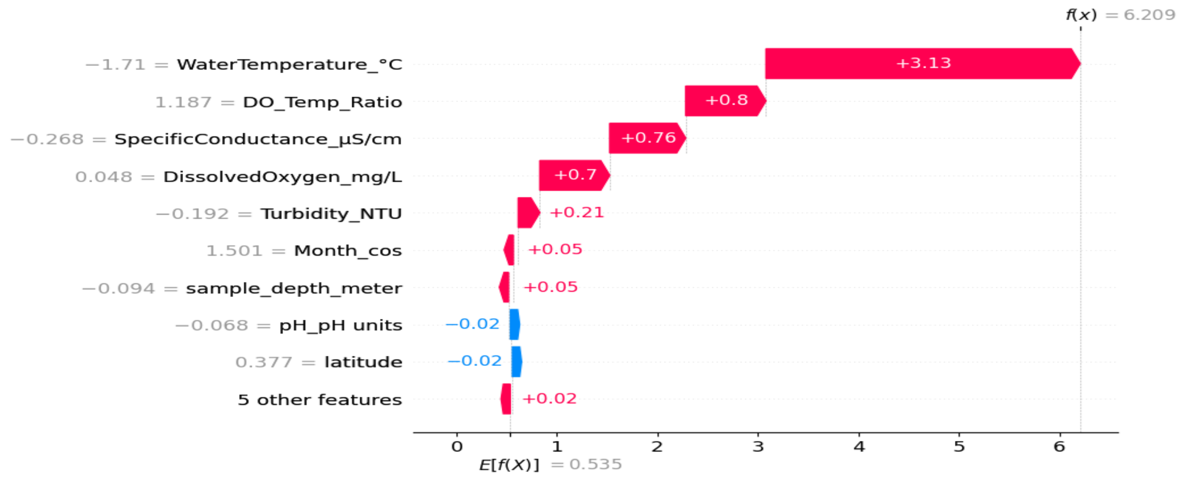




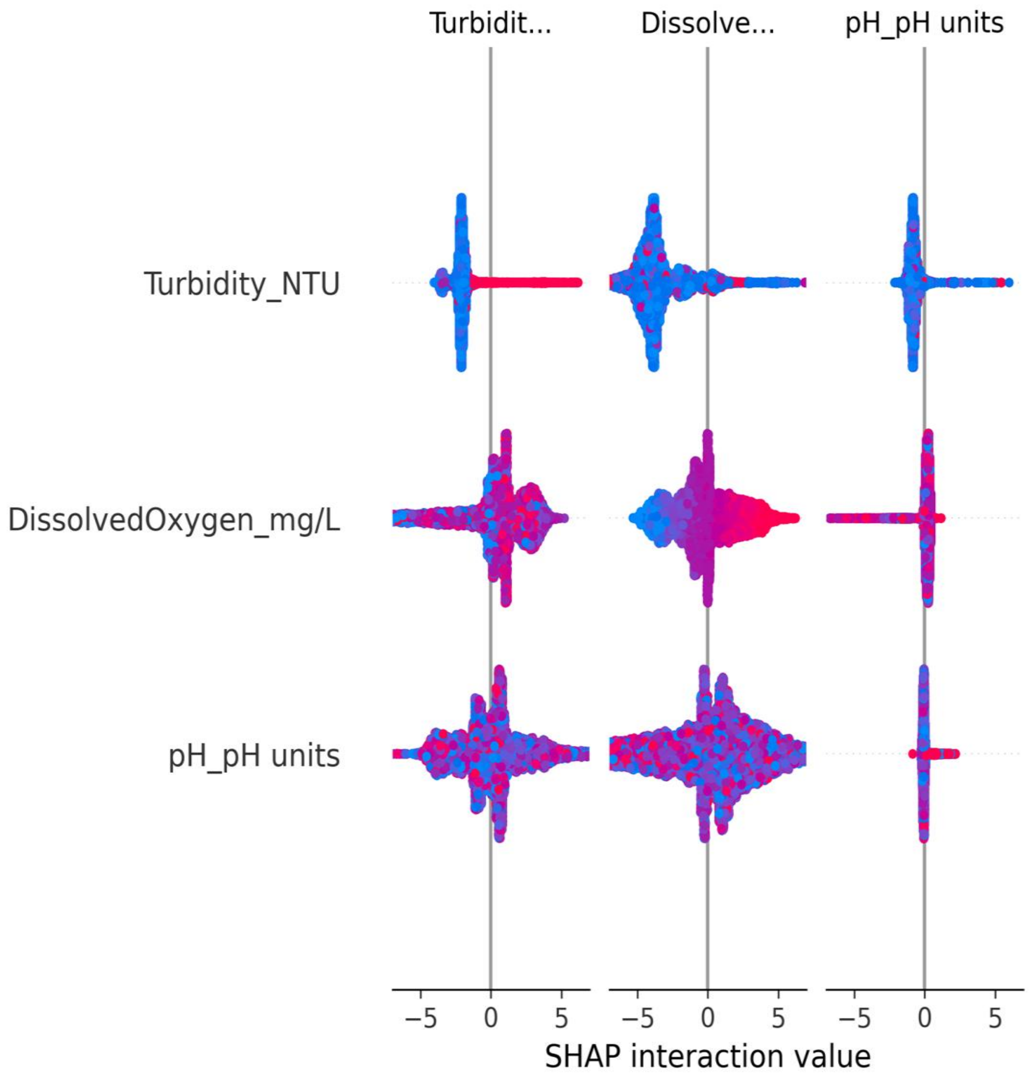
Feature Importance



Top 15 features ranked by XGBoost importance for water quality prediction.



SHAP explanation for a single sample



SHAP summary plot showing global feature importance and effect direction

Demo

Technical Challenges

LARGE DATASET (1.2M ROWS)

MIXED UNITS & NOISY SENSOR READINGS

IMBALANCED CLASSES

SHAP COMPUTATION EXPENSIVE

COORDINATE INCONSISTENCIES

Lessons Learned

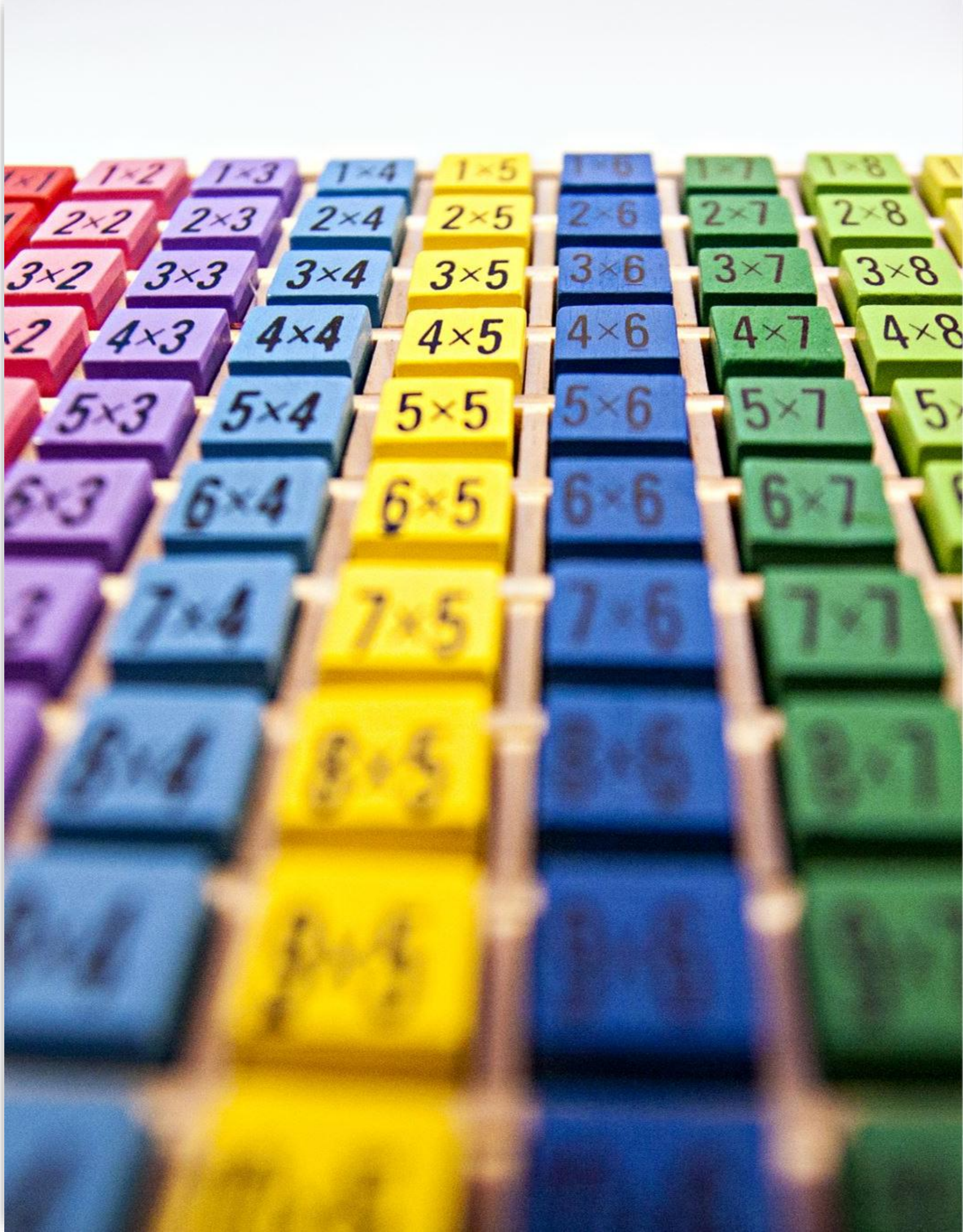
- **Robust Data Engineering**: Unit standardization and outlier filtering were essential for reliable predictions.
- **Hierarchical Imputation**: County-level medians preserved spatial context better than global-only strategies.
- **Feature Selection by Coverage**: 80% missing-value threshold ensured data-driven, informative feature inclusion.
- **Focused Feature Set**: Using five core physicochemical parameters improved interpretability and performance.
- **Temporal Filtering**: Limiting data to 2000–2025 enhanced consistency and removed legacy noise.
- **Scalable Standardization**: Uniform formats enabled seamless integration of new stations without retraining.
- **Class Imbalance Handling**: SMOTE + class weighting ensured fair treatment of minority “Poor” samples.
- **Domain-Informed Features**: DO–Temp Ratio and EPA/WHO thresholds embedded scientific realism.
- **Unified Pipeline Design**: Integrated preprocessing and SMOTE prevented data leakage during evaluation.

Future Work

- Time-series forecasting (DO, pH)
- SHAP explainability optimization
- Adding rainfall, land use, pollution data
- Deep learning extensions (later courses)
- County-level dashboards (GIS)

Conclusion

- ML successfully automates WQI classification
- XGBoost outperforms traditional models
- Dashboard provides actionable insights
- Map visualizations highlight polluted regions
- Supports SDG 6 & environmental protection



Thank You!