Water Safety

Advanced Monitoring & Potability Analytics

Ensuring Sustainable Access to Clean Water through Machine Learning and Predictive Intelligence

A Data Science Project by Valerio Quaranta

Navigating the Complexity of Potability

Turning raw environmental data into actionable insights to ensure safe water for everyone, everywhere.

01

The Chemical Equilibrium

Water safety is not determined by a single factor but by the **delicate synergy** of multiple physical and chemical parameters (pH, Sulfates, Solids, etc.).

02

The Risk of Misclassification

Undetected contamination poses severe risks to the population.
Unnecessary treatments result in significant financial and resource waste for public and private entities.

03

The Mission

Leveraging **Predictive Intelligence** to provide reliable, data-backed decisions for sustainable water management.

Driving Value Through Data Science

Core Objectives

Real-Time Predictive Monitoring:

- Transitioning from slow, traditional laboratory testing to instantaneous potability classification.
- Enabling rapid response to environmental contamination events.

Resource & Cost Optimization:

- Utilizing Feature Selection to identify the most informative chemical sensors.
- Reducing operational expenses by eliminating redundant or low-impact monitoring variables.

Maximum Decision Reliability:

- Ensuring high-stakes accuracy where Safety First is the guiding principle.
- Providing a calibrated risk assessment for public and private stakeholders to act with absolute confidence.

The Chemical Signature: Dataset Characterization

pH	0 – 14	Acid-base balance
Hardness	mg/L	Concentration of Calcium and Magnesium salts
Solids	ppm	Total Dissolved Solids (TDS)
Chloramines	ppm	Primary disinfectants used in water treatment
Sulfate	mg/L	Naturally occurring minerals
Conductivity	μS/cm	Ability of water to conduct electricity
Organic Carbon	ppm	Indicator of purity
Trihalomethanes	μg/L	Byproducts of the chlorination
Turbidity	NTU	Water clarity and suspended particles

Variable Type:

100% Numerical (Continuous independent variables)

Target:

Binary Classification (0: Non-Potable | 1: Potable)

Initial Observation:

Robust but noisy dataset featuring significant missing values; Target **Class Imbalance**

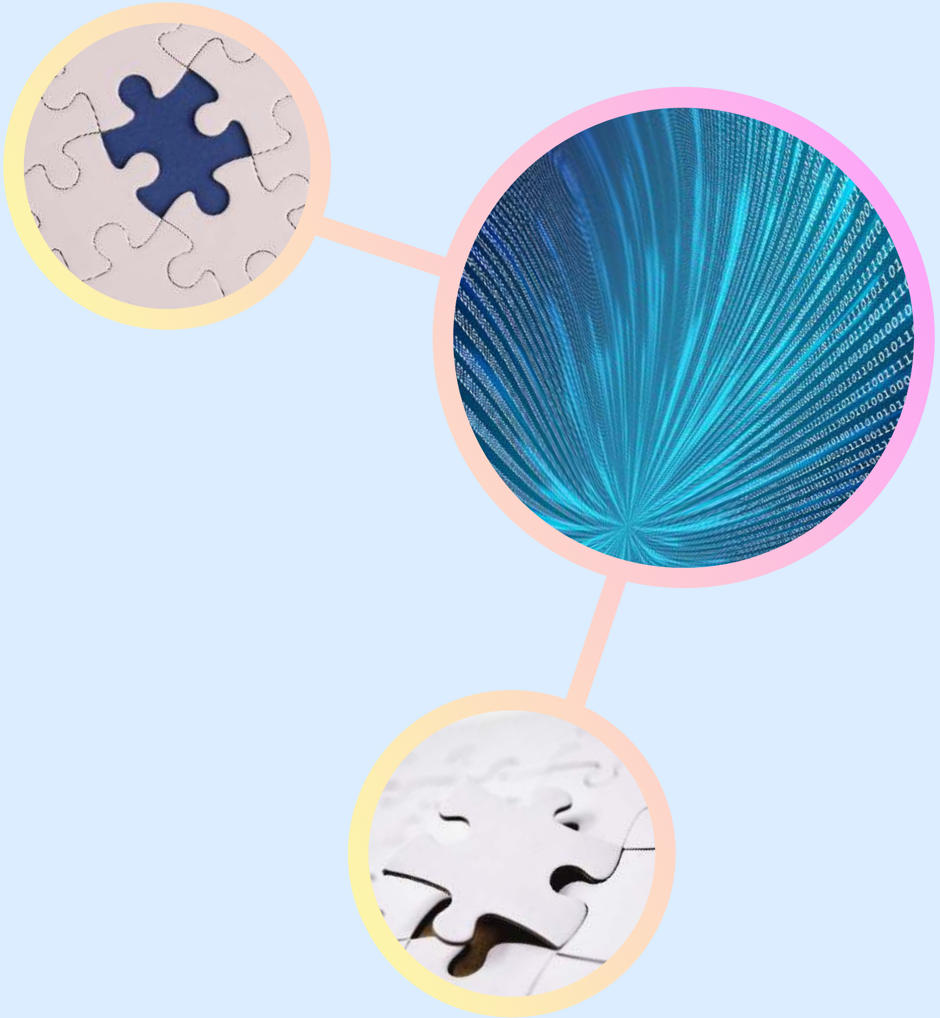
Data Integrity:
The Science of Missing Information

p-value
0.877

Diagnostic Phase

Significant missing values were identified in key parameters. **NaN Correlation**, **Matrix** and **Little's Test** were applied to determine the nature of these gaps.

Conclusion: Since $p > 0.05$, it was confirmed that the data is **Missing Completely At Random (MCAR)**. This mathematically proves that the absences are accidental and not biased by external factors.



Tactical Phase

Unlike simple median filling, the **Iterative Imputer** treats every missing variable as a function of the others.

In environmental science, parameters are rarely independent. This method preserves the **multivariate correlations** between minerals and chemicals, ensuring a much more realistic dataset for the AI.

Cleaning the Signal: Robust Anomaly Detection

Beyond Simple Thresholds

- Environmental data often features extreme values. Standard methods like IQR (Interquartile Range) only look at variables in isolation, potentially missing complex relational anomalies.
- **The Risk:** Deleting too much data can blind the model to rare but real environmental events, while leaving too much noise can degrade predictive accuracy.

The Solution: Isolation Forest (ISO)

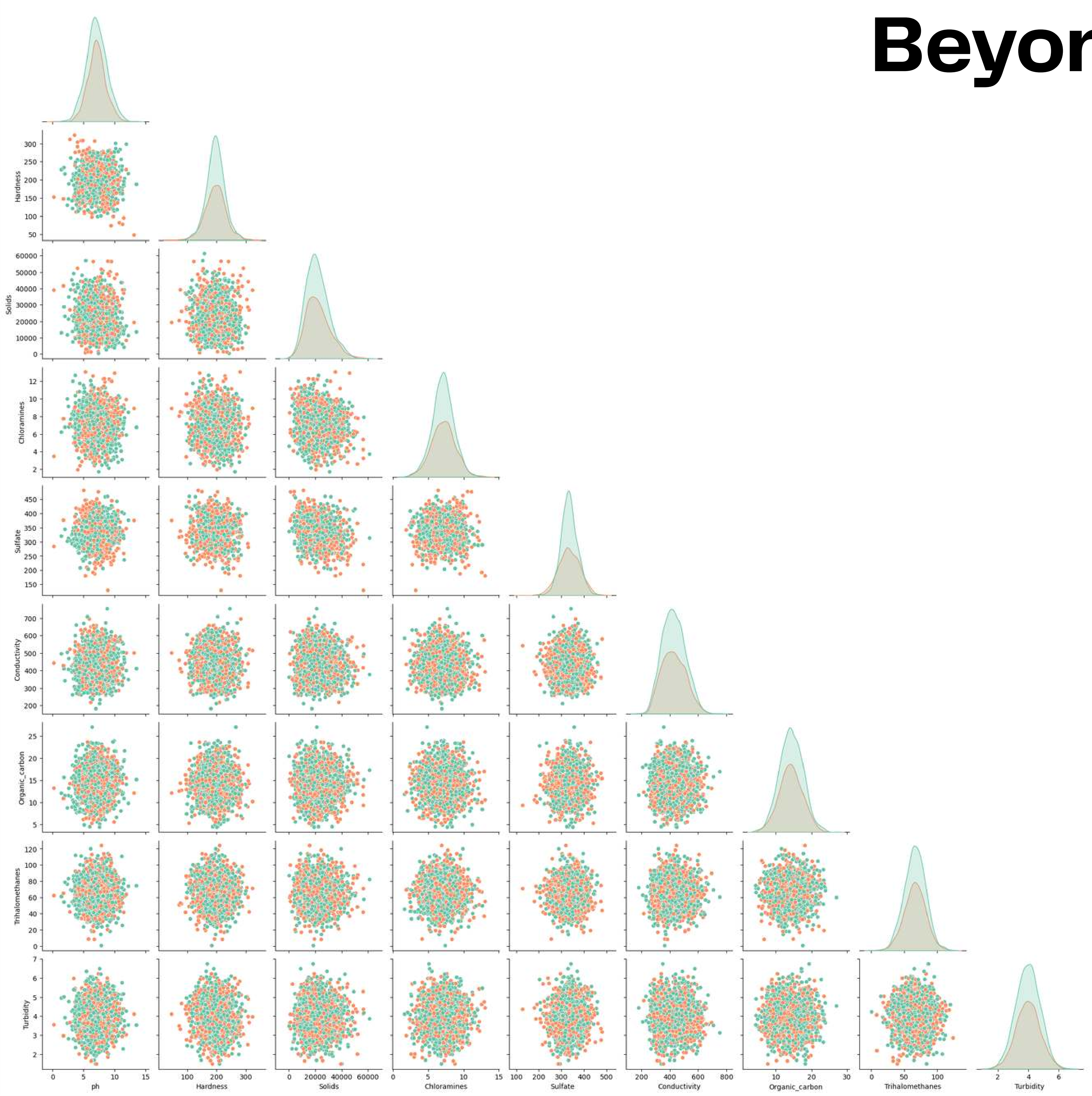
An Isolation Forest algorithm was implemented. Instead of checking if a value is "too high" or "too low," ISO identifies anomalies by how easily they can be "isolated" within a **multidimensional tree structure**.

By setting a **contamination** threshold of 0.05, we focused only on the most extreme 5% of samples—those most likely to be measurement errors or irrelevant noise.

Anomalies Removed: 123 extreme samples

This surgical cleaning preserved the core chemical signal while protecting the training phase from being distorted by non-representative data points.

Beyond Linear Correlations



Multivariate analysis revealed an exceptionally **weak linear correlation** between physical-chemical parameters and water potability.

Hidden Non-Linear Patterns

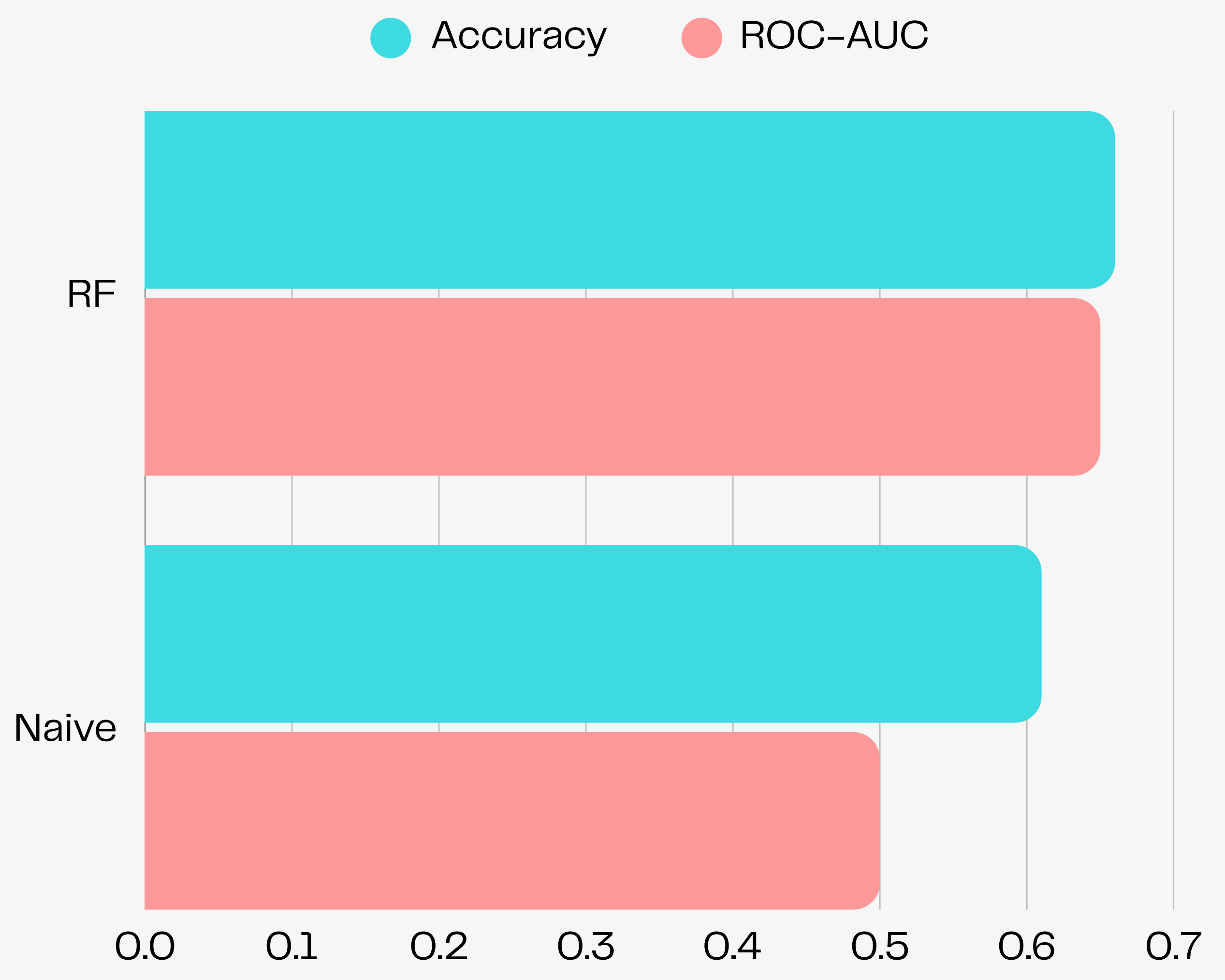
The lack of linear correlation is not a lack of information.
The data hides complex, non-linear dependencies.
Potability isn't determined by one factor, but by the **specific intersection** of multiple chemical thresholds.

Setting the Bar: The Baseline Evaluation

The Naive Baseline

If we simply predicted Non-Potable for every sample based on the majority class (61%), we would achieve **61% Accuracy**

The Limitation: While the accuracy seems acceptable, this naive approach has **zero predictive power** (0% Precision/Recall for potable water) and provides no utility for safety monitoring.



The Main Baseline: Random Forest

A Random Forest Classifier requires minimal preprocessing and handles non-linear data naturally.

The Result:

Accuracy: 66%

Roc-Auc: 0.655

This confirms that there is a significant learnable signal in the data. The model is already 5% better than a blind guess and is beginning to distinguish between classes.

.The Baseline is our floor, not our ceiling. It confirms that the chemical parameters hold the key to predicting potability, justifying the move toward more advanced models.

Strategic Modeling

To determine the most efficient way to monitor water quality, two parallel modeling strategies were developed

Strategy A
Full Features

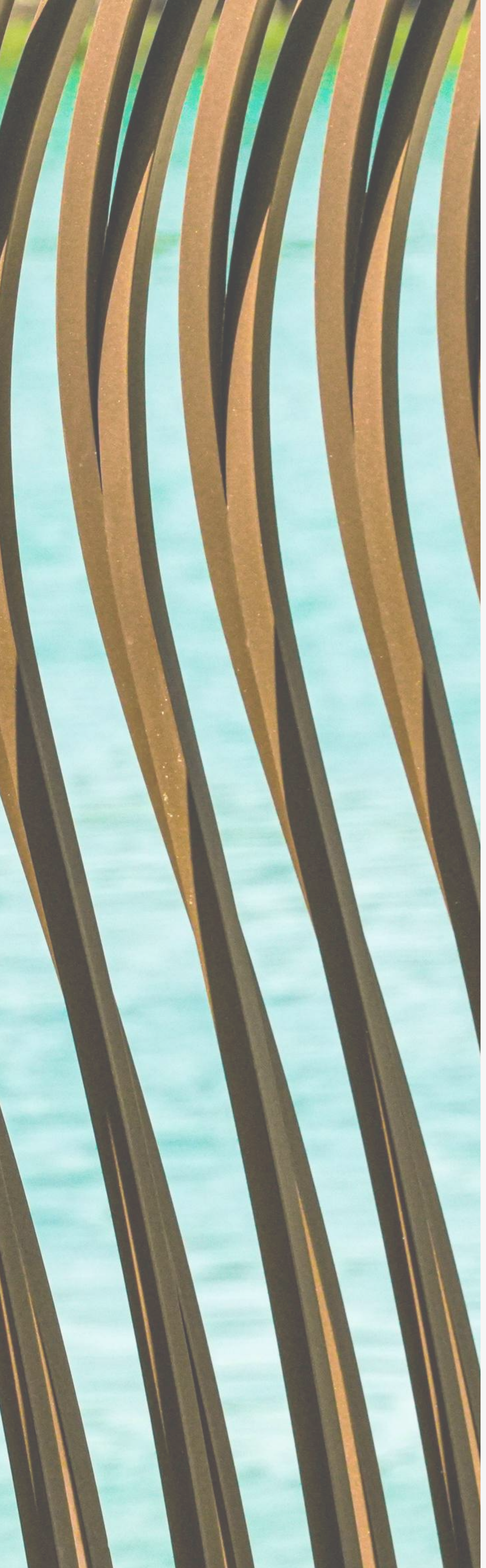
Utilizing the complete array of 9 chemical and physical parameters to capture every possible nuance.

Strategy B
Reduced Features

Excluding variables identified as noise or low-impact by EDA, T-Tests and Baseline Feature Importance.

Why Feature Selection?

- **Sensor Optimization:** Reducing the number of required parameters directly lowers hardware and maintenance costs for water monitoring stations.
- **Noise Mitigation:** Removing non-significant variables like ***Turbidity, Conductivity*** and ***Trihalomethanes*** helps models focus on the core drivers of potability, potentially improving generalization.



Selecting the Engines: SVC and XGBoost

Support Vector Classifier (SVC)

The Logic: SVC uses the "**Kernel Trick**" to project water data into higher dimensions, finding the optimal boundary (hyperplane) where a linear split is impossible.

The Goal: Capturing the subtle, non-linear chemical interactions that define safe water. It is exceptionally robust for medium-sized, high-quality datasets.

XGBoost (Extreme Gradient Boosting)

The Logic: An ensemble method that builds a sequence of decision trees, where each new tree corrects the errors of the previous ones.

The Goal: Handling complex feature interactions and potential noise with extreme efficiency. It is the gold standard for high-performance predictive modeling in data science.

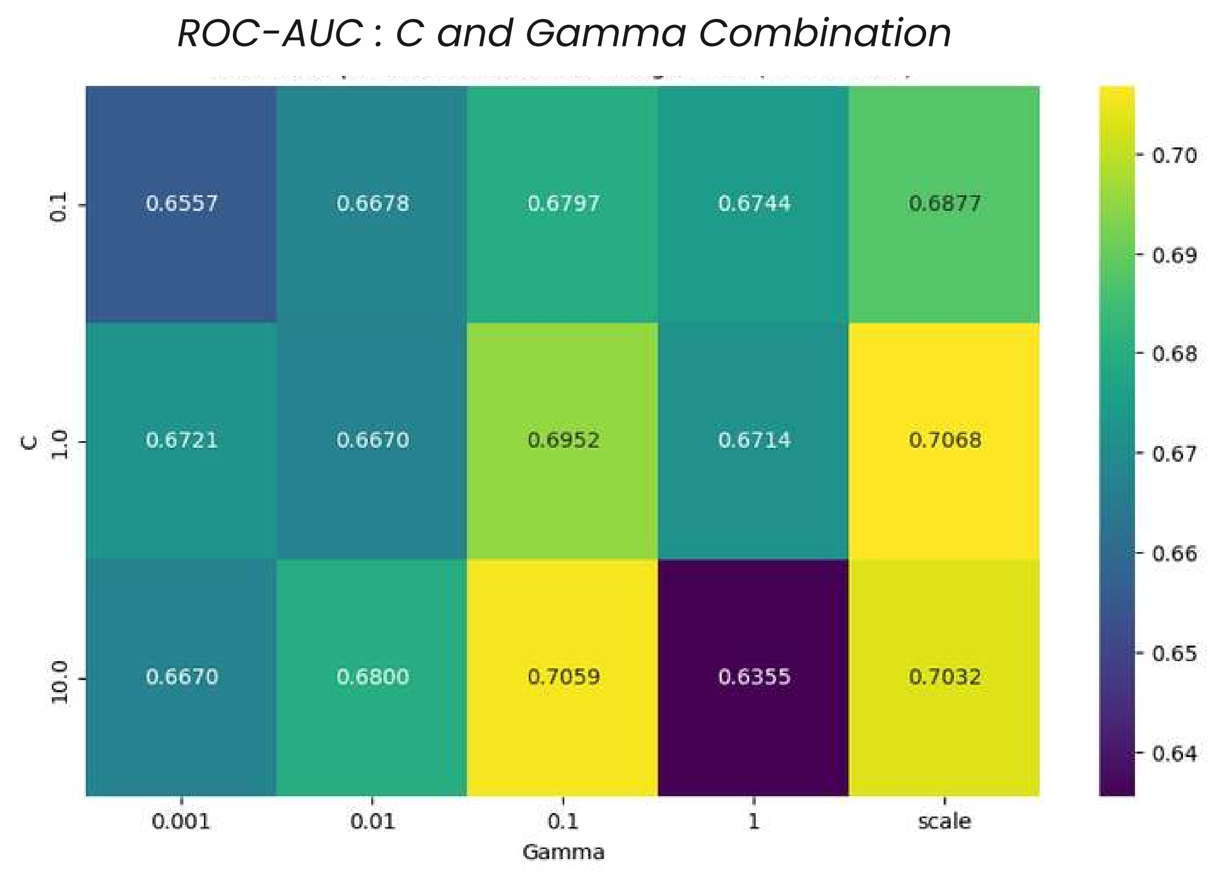
Hyperparameter Optimization & Risk Control

The Tuning Strategy: Grid Search CV

Cross-Validation: Every configuration was tested across multiple data folds (Stratified K-Fold) to ensure the results were stable and not a matter of luck.
Multiphase Approach: We moved from a broad Logarithmic Scale (to find the order of magnitude) to a Refined Linear Scale (to pinpoint the exact optimal values).

SVC Calibration

Balancing **C (Margin)** and **Gamma (Influence)** to find a boundary that is flexible enough for chemistry but rigid enough to ignore noise.



The XGBoost Memory Trap

Initial tests showed the model was memorizing the training data. We solved this by:

Regularization: Introducing **L2 Penalty (Lambda)** to punish excessive complexity.
Stochasticity: Implementing **Subsampling** (40%) to force the model to learn from random subsets of data.

Outcome

Successfully reduced the gap between training and validation performance.

Performance Benchmark

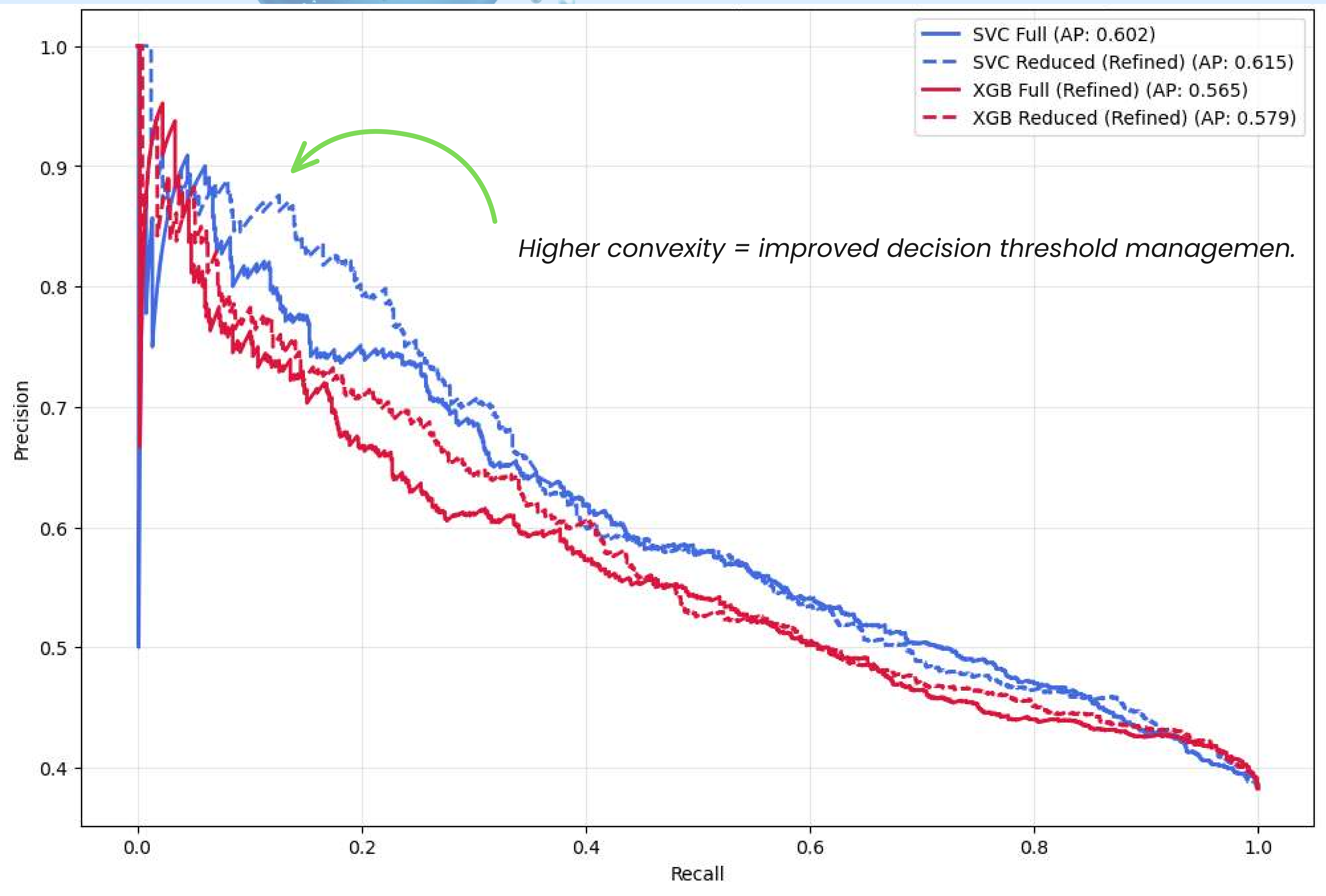
The Power of Selection

The "Reduced" models consistently match or exceed the "Full" models.

Efficiency

We have proven that we can achieve superior results with **3 fewer sensors**, optimizing both computational speed and monitoring costs.

Reliability Under Pressure: The PR Curve Analysis



Beyond Accuracy: Why Precision-Recall Matters

01

- In water monitoring, classes are often imbalanced. Accuracy can be deceiving.
- The Precision-Recall (PR) Curve evaluates how well the model identifies potable water (**Recall**) without triggering too many false alarms (**Precision**).

The Impact of Feature Selection

02

- Increased Precision: The "Reduced Refined" strategy didn't just simplify the model; it increased the **Average Precision (AP)**.
- **Noise Reduction**: By removing redundant variables, we reduced the confusion of the model, allowing for a cleaner separation between safe and unsafe samples.

Safety-First Thresholds

03

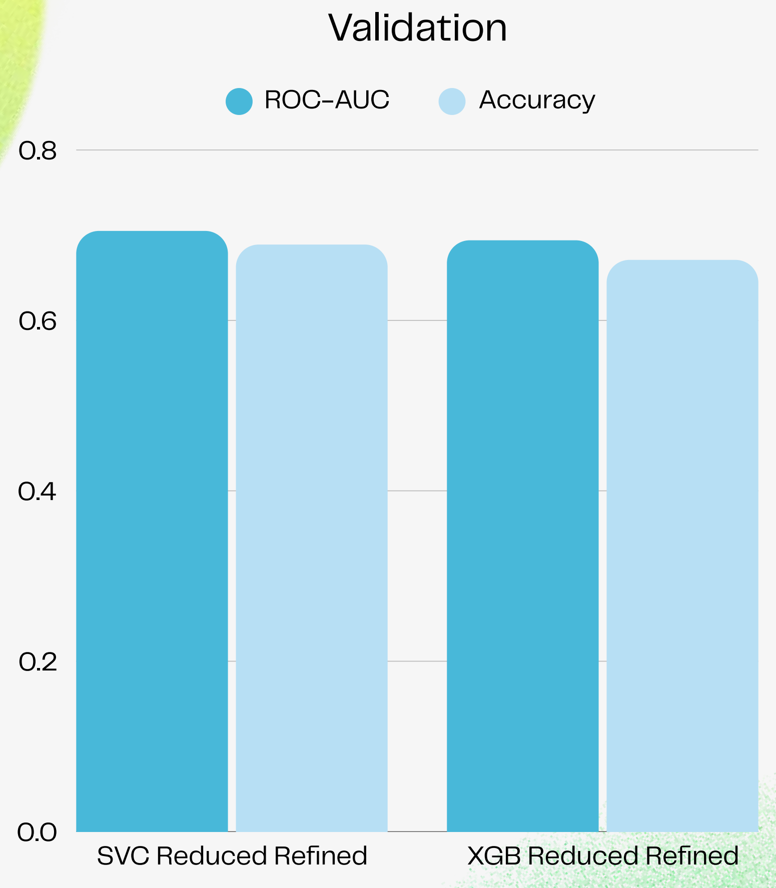
- The curve demonstrates that SVC maintains high precision even as we increase our sensitivity.
- This ensures that when the model labels water as Potable, the confidence level is structurally robust, supporting the company's commitment to **Public Health Safety**.

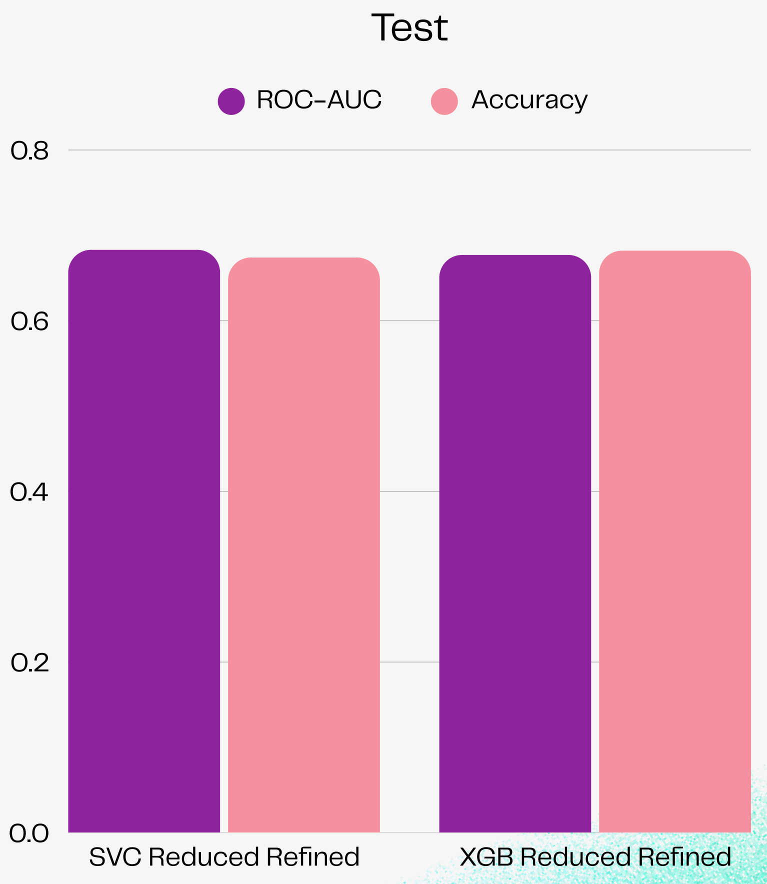
Final Validation:
Performance on Unseen Data

Transitioning from training to the Final Test Set, the models maintained **remarkable stability**.
The hierarchy established during cross-validation remained consistent, proving that our selection process was statistically sound.

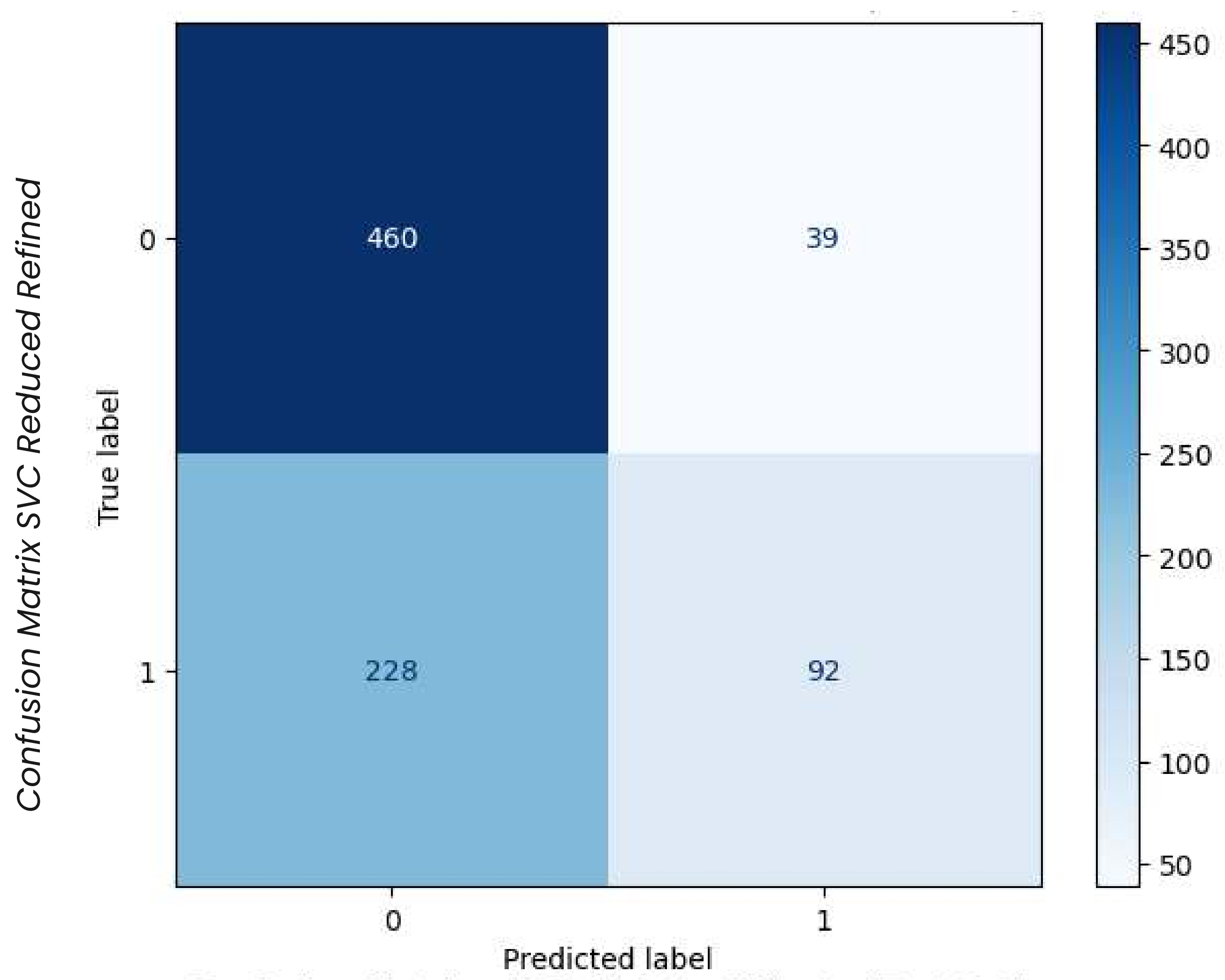
SVC Reduced Refined (Best Stability) successfully achieved a **ROC-AUC of 0.683**. It remains the most balanced model for probabilistic ranking.
XGBoost Reduced Refined (Highest Accuracy) reached **68.25% Accuracy**, showing a slightly more decisive classification behavior in final testing.

Both champions significantly outperformed the **Random Forest Baseline** (Accuracy: 66.0%, ROC-AUC: 0.655).
This delta represents the Value Added by our advanced feature engineering and hyperparameter tuning phases.





Error Analysis: Prioritizing Public Safety



The "Safety-First" Bias

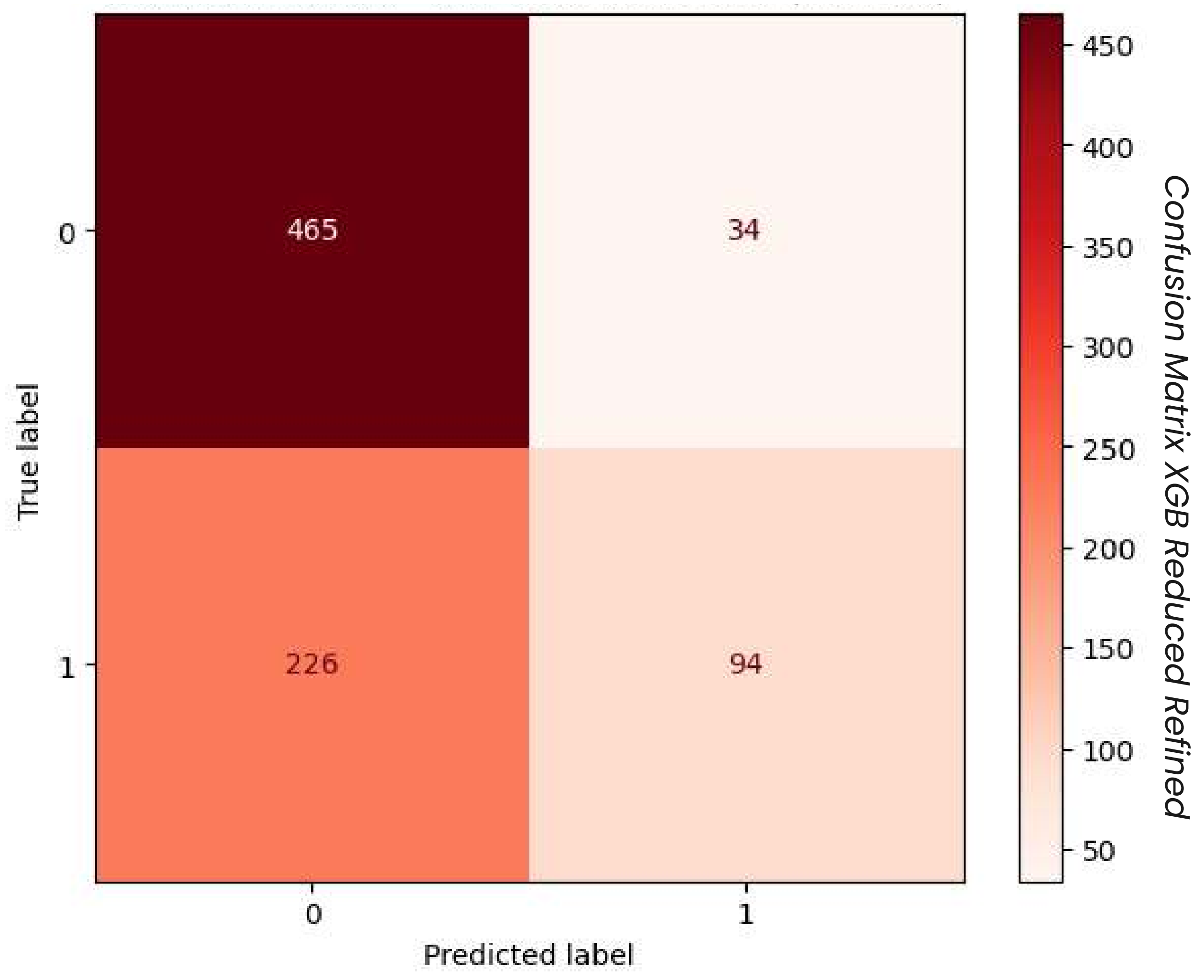
The models exhibit a markedly **conservative behavior**. All classifiers prioritize avoiding a **False Positive** (the most dangerous error in this industry) over maximizing raw accuracy.

- **Result**: A high Recall for the Non-Potable class (>92%), ensuring that contaminated samples are almost never mistakenly classified as safe.

Managing the Trade-off

- **False Positives**: the refined models minimized these to a deeply low level. Between SVC and XGBoost, we can observe only ~35-40 cases out of the entire test set.

- **False Negatives** (Operational Caution): Most errors are False Negatives—labeling safe water as unsafe. While this may cause temporary treatment delays, it guarantees zero compromise on human health.



Quantifying Certainty

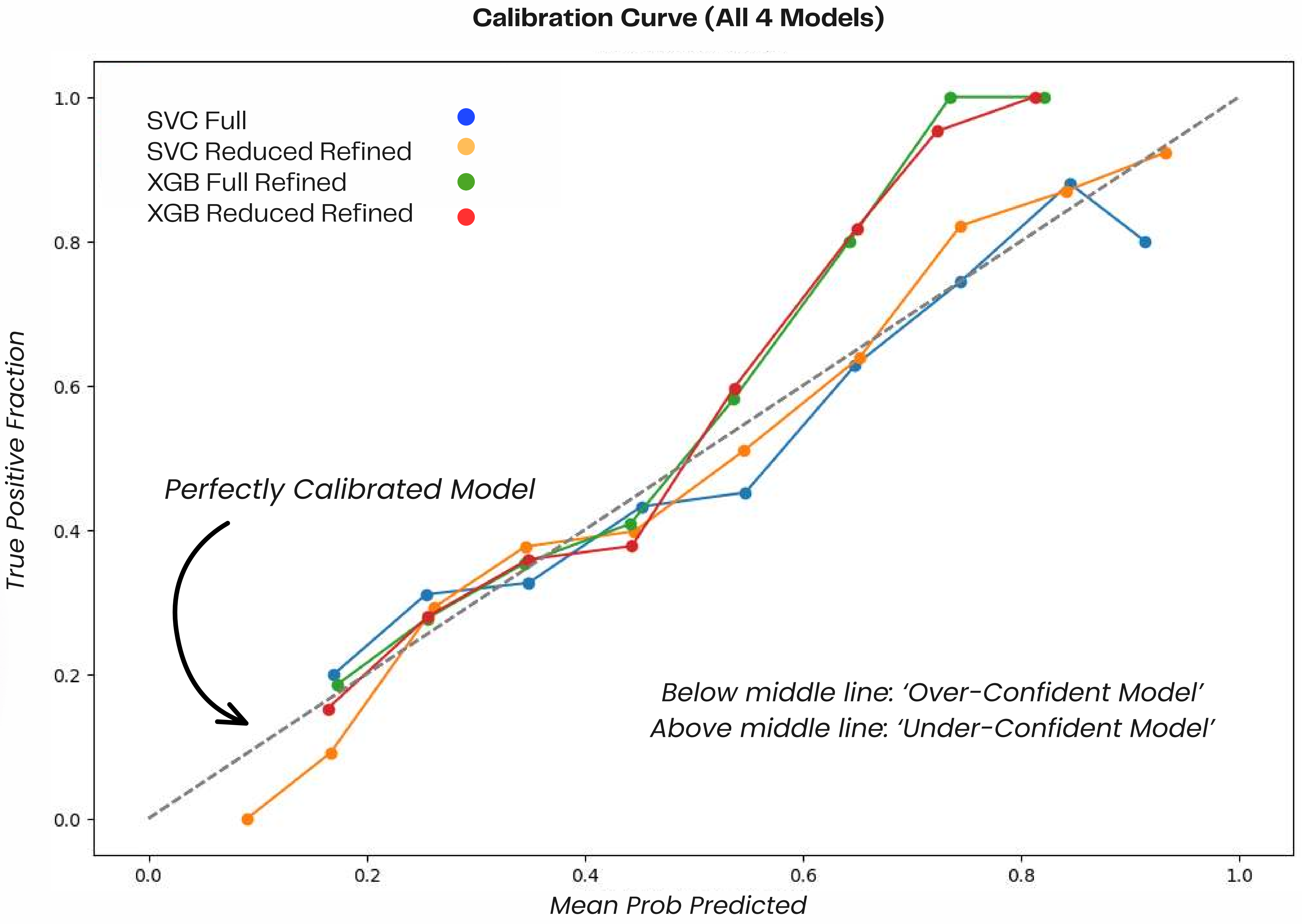
A great model shouldn't just predict; it should provide a **reliable probability**. If the model says there is a 70% chance the water is potable, it should be right exactly 70% of the time.

The **Calibration Curve** visualizes this honesty, showing how closely the model's confidence aligns with real-world outcomes.

Why it matters for Water Management?

If we say **SVC** is the most "calibrated" model, this means its probability estimates are the most trustworthy for decision-makers.

When the system reports a low *Confidence* sample, it allows human operators to intervene, creating a perfect **Human-in-the-Loop** safety system.



The **Brier Score** calculates the mean squared difference between predicted probabilities and the actual outcome (the lowest the best):

- **SVC Reduced Refined (Winner)**: Achieved a score of **0.209**.
- **XGBoost Reduced Refined**: Achieved a score of **0.219**.

Shared Failures

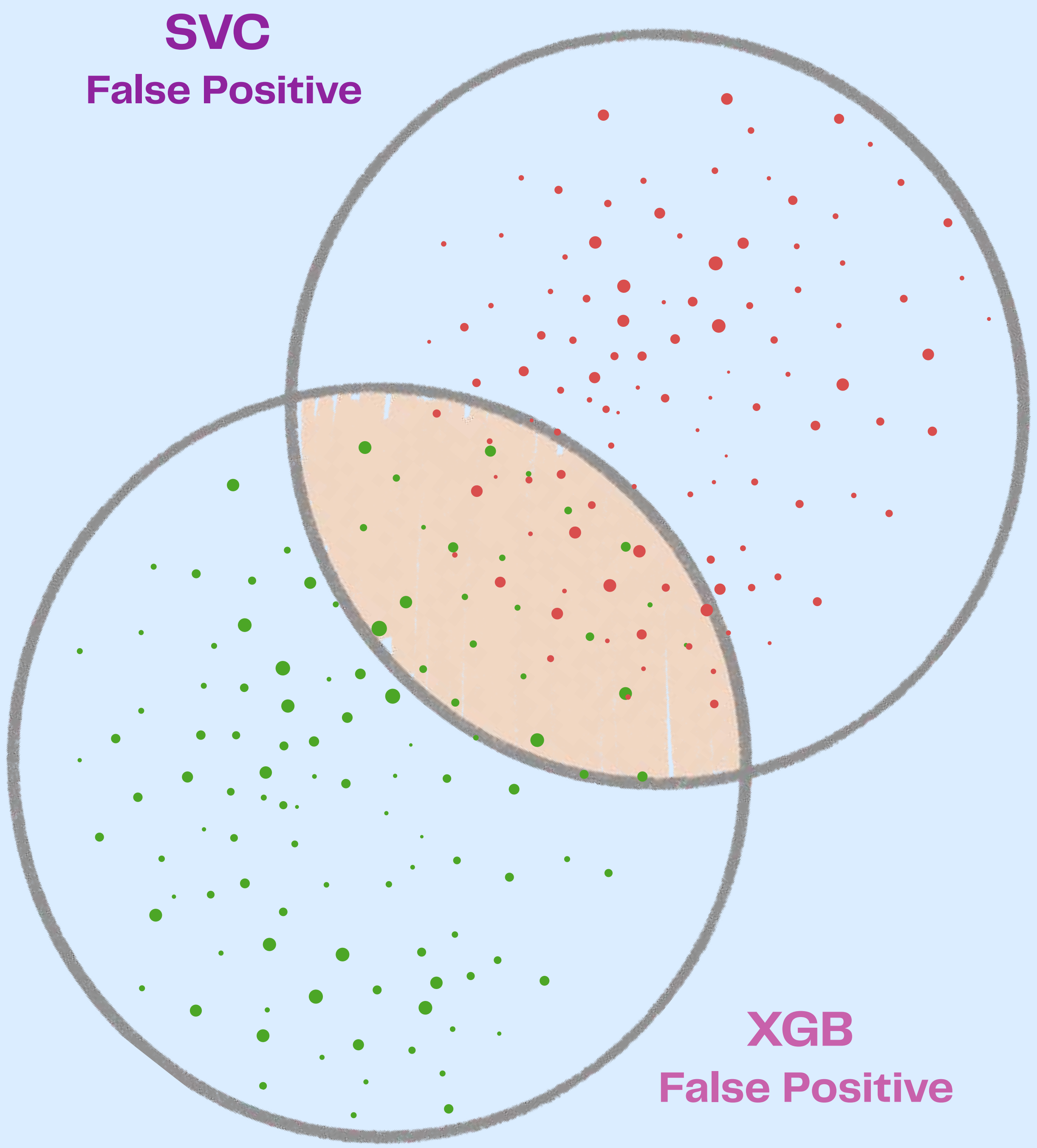
18 specific samples were identified (almost half), that both models (SVC and XGBoost) failed to classify in the exact same way.

- **Conclusion**: These are not random errors. They represent samples with intrinsically ambiguous chemical profiles where classes overlap perfectly. It would be hard to resolve this ambiguity without additional sensors or data sources.

Regional Uncertainty

Errors tend to cluster within specific ranges, such as **near-neutral pH levels** (between 6.5 and 7.5).

- **Actionable Insight**: In these gray zones, the system can be programmed to automatically trigger a request for human validation or supplementary laboratory testing.



Are the errors caused by the missing data which were filled in during preprocessing?

The analysis shows that the count of imputed values in misclassified samples is identical to that of correctly predicted ones. This confirms that the **Iterative Imputation** was impartial

The Winner: Maximum Intelligence, Minimal Footprint

After exhaustive testing, the **Support Vector Classifier on the Reduced Dataset** stands as the definitive solution.

ROC-AUC

0.683

Accuracy

67.4%

- **Noise Resistance**: Unlike XGBoost, which showed signs of sensitivity to environmental noise, SVC maintained a more stable and geometric decision boundary.
- **Generalization**: The minimal gap between validation and test scores ensures that this model is robust enough for diverse geographic locations.

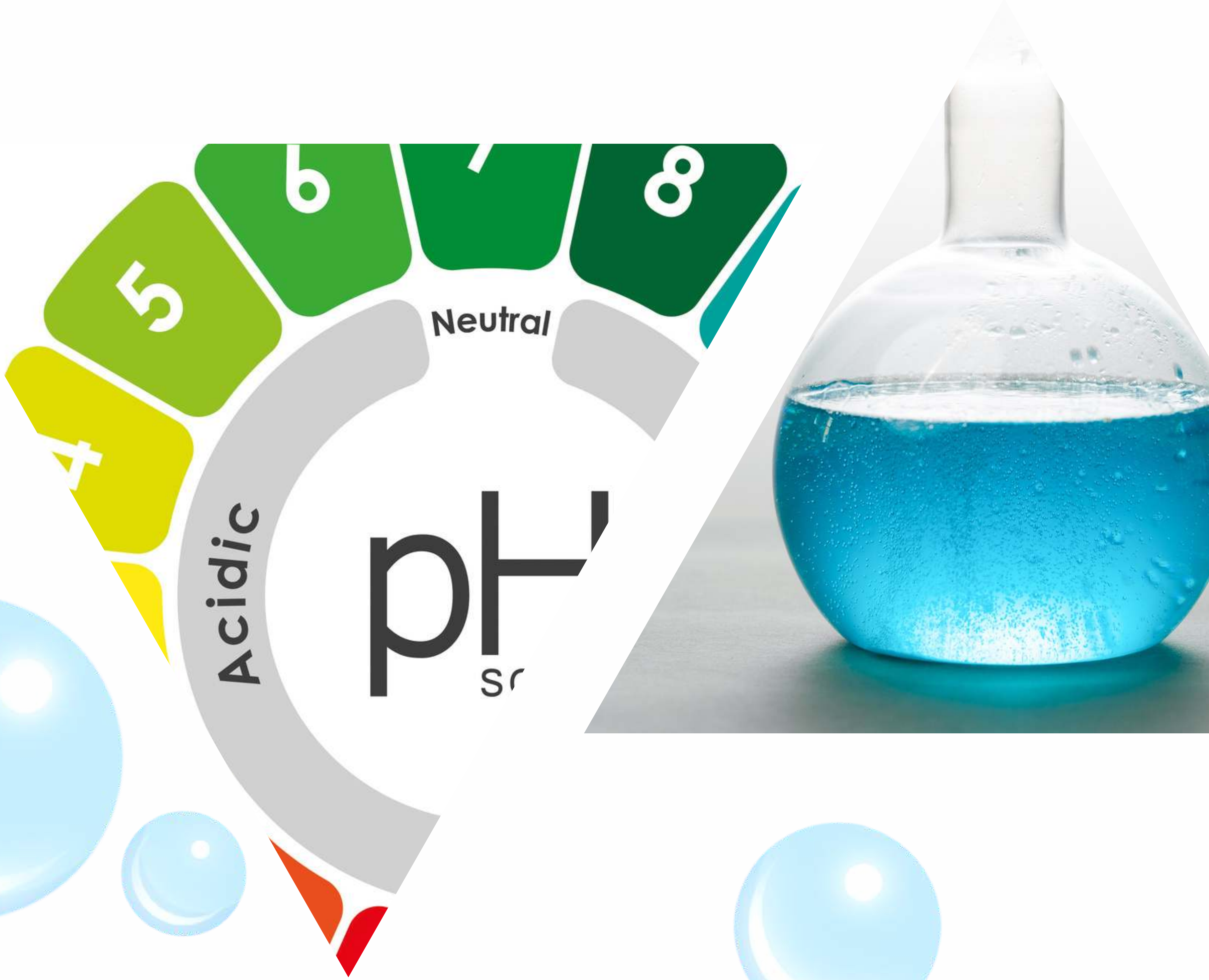
Key Chemical Drivers

The feature importance scores from the XGBoost model consistently identified **pH** and **Sulfate** levels as the most influential predictors in the dataset.

These two variables act as the foundational pillars of water potability, carrying more predictive weight than all other chemical parameters

The Interaction

The model found that the **interaction** between high Sulfates and specific pH levels is a **red flag** for safety, a nuance that simpler models failed to capture.



Deployment Strategy & Future Roadmap

Threshold Moving

- Instead of using a standard 0.5 probability cutoff, implement Threshold Moving to favor extreme caution.
- By shifting the decision boundary, artificially force the model to be more skeptical of potability, further reducing the risk of False Positives to near-zero levels.

Ensemble Modeling

- Deploy a Soft Voting Ensemble that combines the geometric precision of SVC with the gradient-boosted logic of XGBoost.
- By averaging the predicted probabilities from both champions, produce a more robust output that balances out the individual biases of each algorithm.

Automated Alerting

- Define a Safe Buffer (e.g., 40%–60% probability) where the model's confidence is statistically thin.
- Any sample falling into this zone triggers an immediate Real-Time Alert. These samples are flagged for mandatory human verification or high-priority laboratory re-testing, ensuring no ambiguous case ever reaches the consumer.

Thank You

Dataset Source: *Water Potability Dataset*

Project Repository: *Water Safety* (GitHub)

Contacts: *vlr.quaranta@gmail.com; LinkedIn*