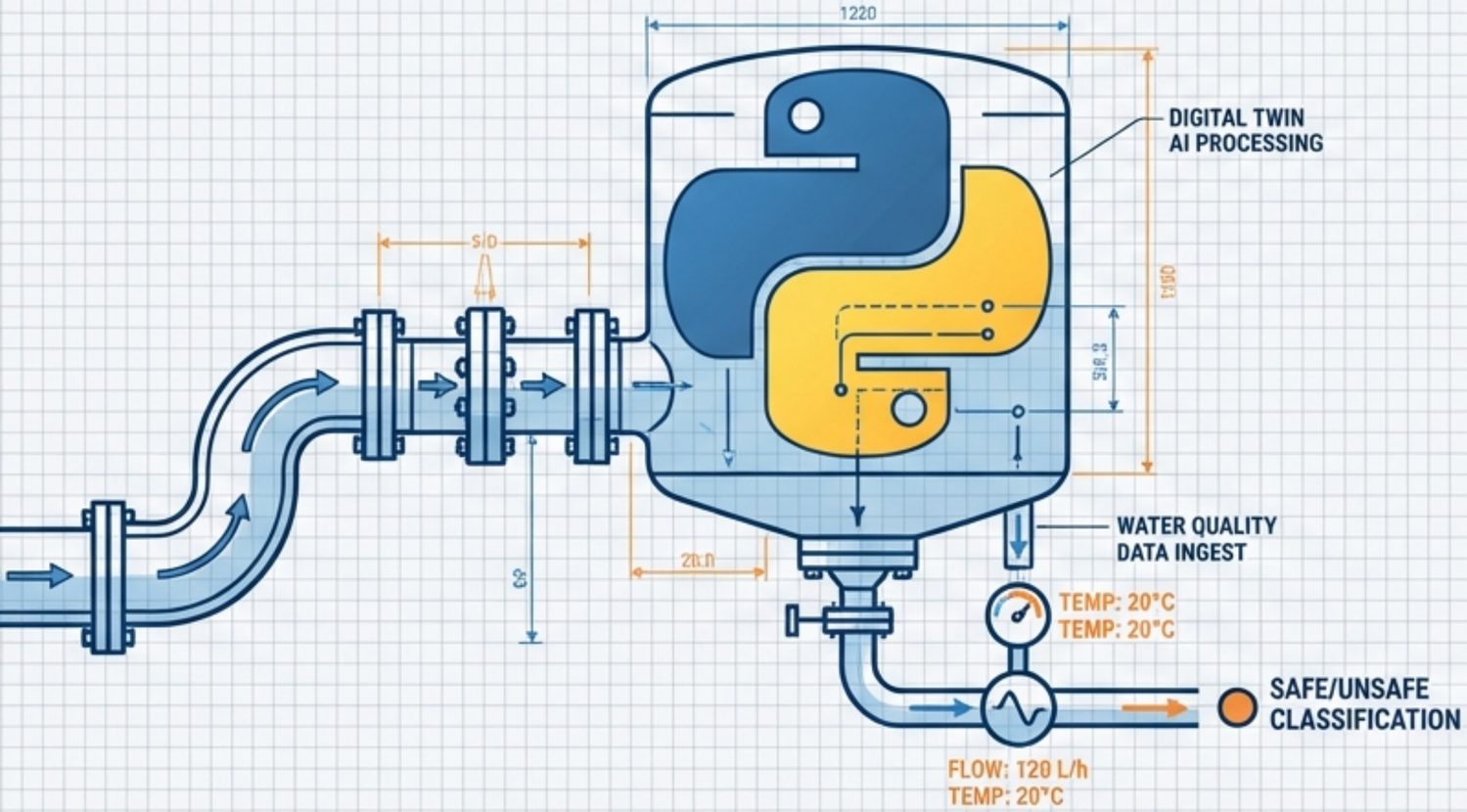
Unit 12: 化工 AI 應用——水質安全分類預測系統

基於機器學習的飲用水可飲用性自動化評估



SPEC SHEET

日期: 2026-01-28

版本: v1.0

任務: 二元分類 (Binary Classification)

核心指標: 召回率 (Recall) > 95%

數據集: 3,276 樣本 (9 特徵)

專案背景：危機與轉機

當前危機 (Current Crisis)

全球水質危機

根據 WHO 統計，全球 **20 億人**使用受污染水源。

每年 **50 萬人**死於水源性疾病。

- **耗時**：24-48 小時
- **昂貴**：NT$ 5,000-10,000/次
- **滯後**：無法即時攔截

System Upgrade

數位解方 (Digital Solution)

機器學習數位防線

利用 **9** 項水質數據構建即時預測模型。

- **即時預測**：< 1 秒
- **低成本**：僅需感測器數據
- **全時監控**：24/7 自動化

輸入參數規格：9 項關鍵水質指標

pH (酸鹼值)

- 範圍：0 – 14
- WHO 標準：6.5 – 8.5

Hardness (硬度)

- 單位：mg/L
- 軟水標準：< 60 mg/L

Solids (TDS - 總固體溶解量)

- 單位：ppm
- WHO 標準：< 1000 ppm
- 備註：關鍵污染指標

Chloramines (氯胺)

- 單位：ppm
- EPA 標準：≦ 4.0 ppm

Sulfate (硫酸鹽)

- 單位：mg/L
- WHO 指導值：< 500 mg/L

Conductivity (導電度)

- 單位：µS/cm
- 與 TDS 高度相關

Organic_carbon (TOC)

- 單位：ppm
- EPA 標準：< 2 ppm

Trihalomethanes (THMs)

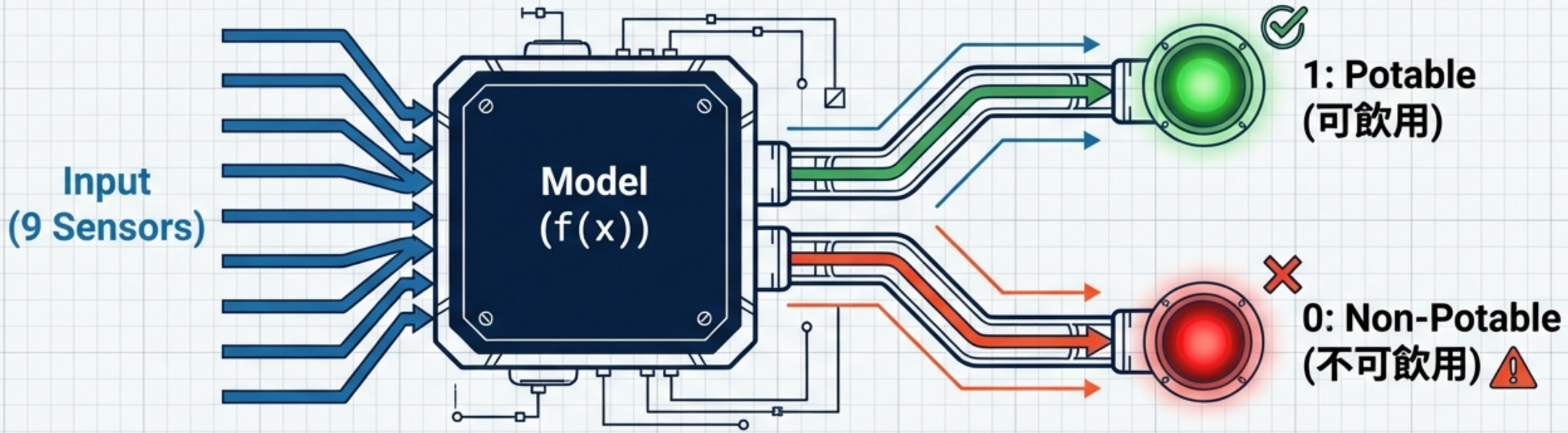
- 單位：µg/L
- 致癌副產物
- EPA 標準：≦ 80 µg/L

Turbidity (濁度)

- 單位：NTU
- WHO 標準：< 5 NTU

- Data Standards: WHO Guidelines & US EPA Regulations

系統邏輯定義：二元分類閘門



系統性能目標 (System Specs):	
Recall (召回率):	> 95% (Critical Target)
AUC (區分力):	> 0.90
Latency (延遲):	< 1 sec

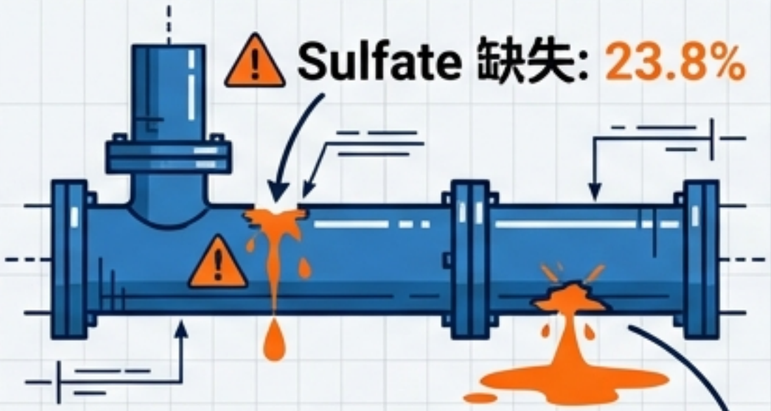
CRITICAL SAFETY NOTE: False Negative is the Enemy. (漏放不安全的水是最危險的錯誤)

原料檢驗：數據探索性分析 (EDA)

數據概況 (Data Overview)

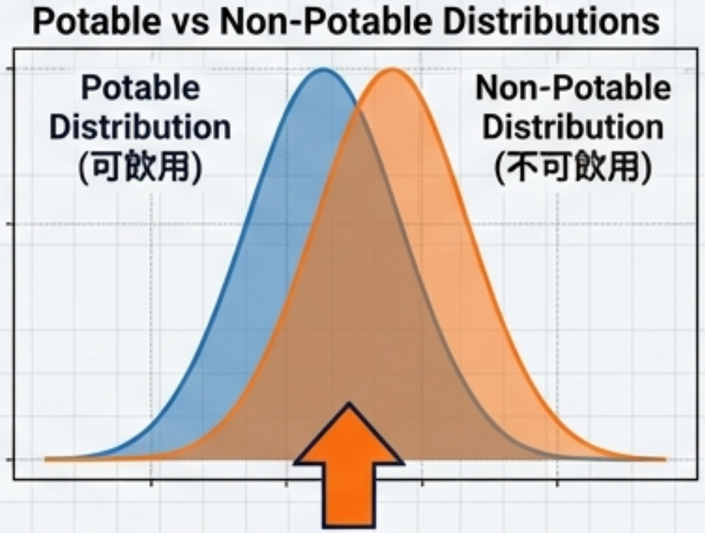
- 總樣本數: 3,276 筆
- 類別比例: 61% (0) vs 39% (1)
- 狀態: 輕度不平衡 (Mild Imbalance)

缺陷檢測 (Defect Detection)



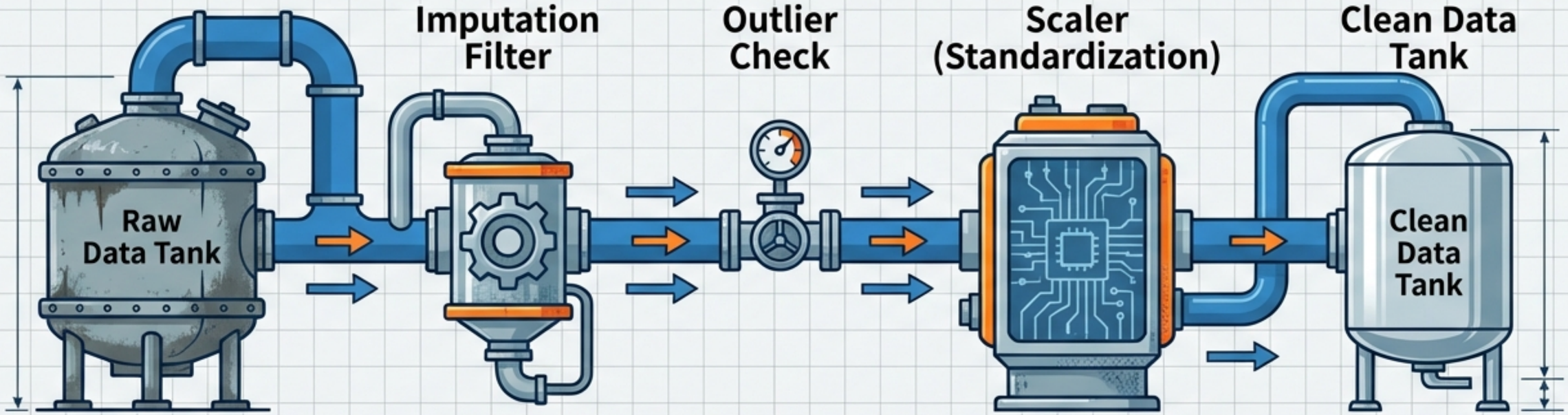
修復策略: 中位數填補 (Median Imputation)

特徵重疊 (Feature Overlap)



特徵高度重疊，無單一線性開關 (High Overlap)

前處理程序：數位過濾系統



Roboto Mono

含缺失值與異常值

(Fills Sulfate/pH holes)

中位數填補 (Median)

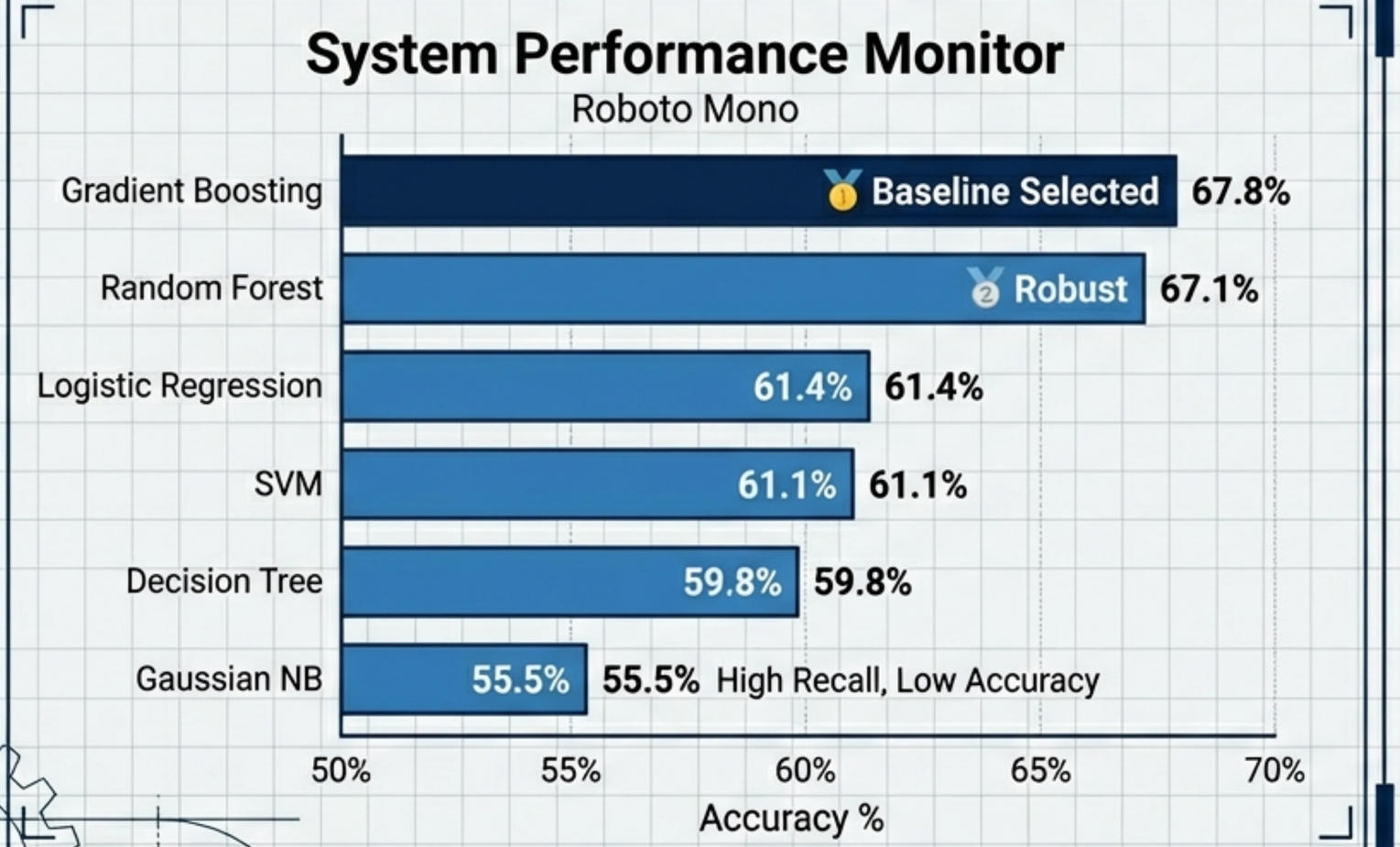
保留化學異常值 (IQR Method)

(Normalizes 0-14 pH vs 60k Solids)

標準化 (Z-score)

Ready for Training

核心引擎測試：模型基準評比



工程分析 (Analysis)：

線性模型 (LR, SVM) 表現不佳，因特徵相關性低 (<0.05)。

樹模型 (GB, RF) 能處理非線性複雜度，獲選為核心引擎。

風險評估：混淆矩陣分析

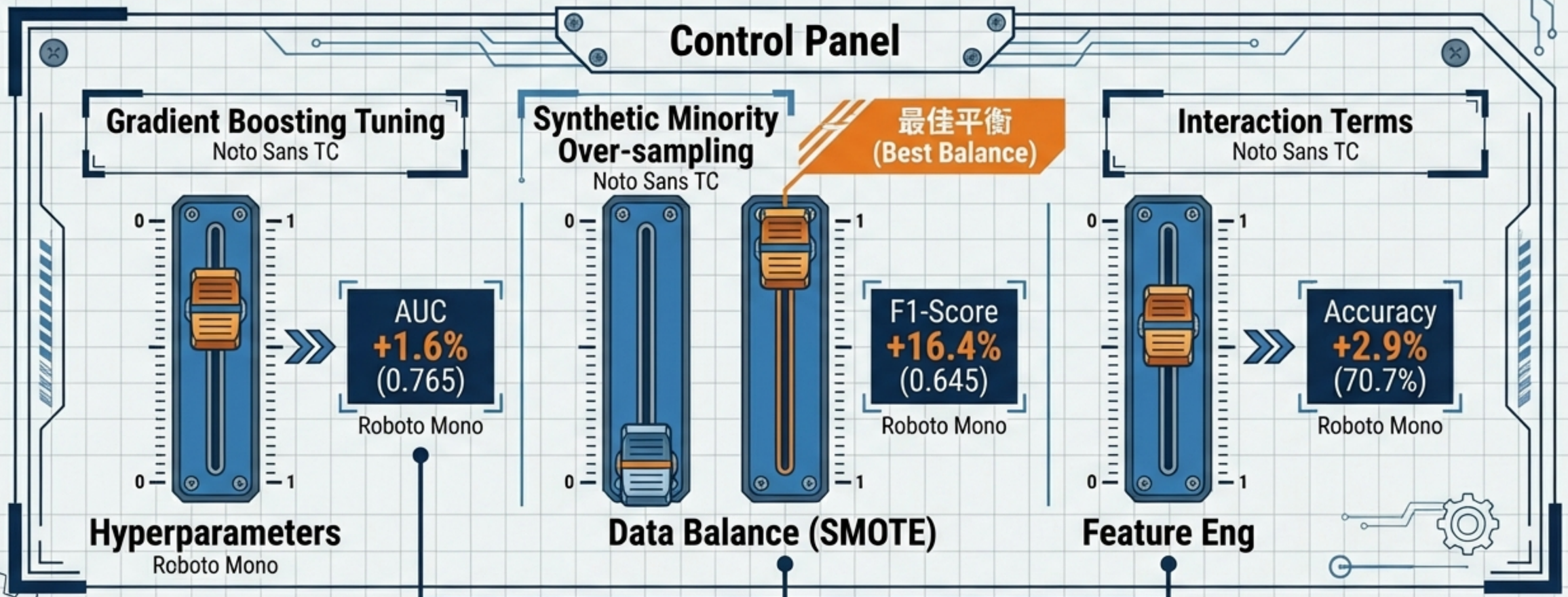
嚴重缺失 (Critical Failure)

漏檢率 (FN Rate)：48.8%

意義：125 個不安全水樣本被誤判為安全。

行動：必須調整閾值 (Threshold Adjustment)。

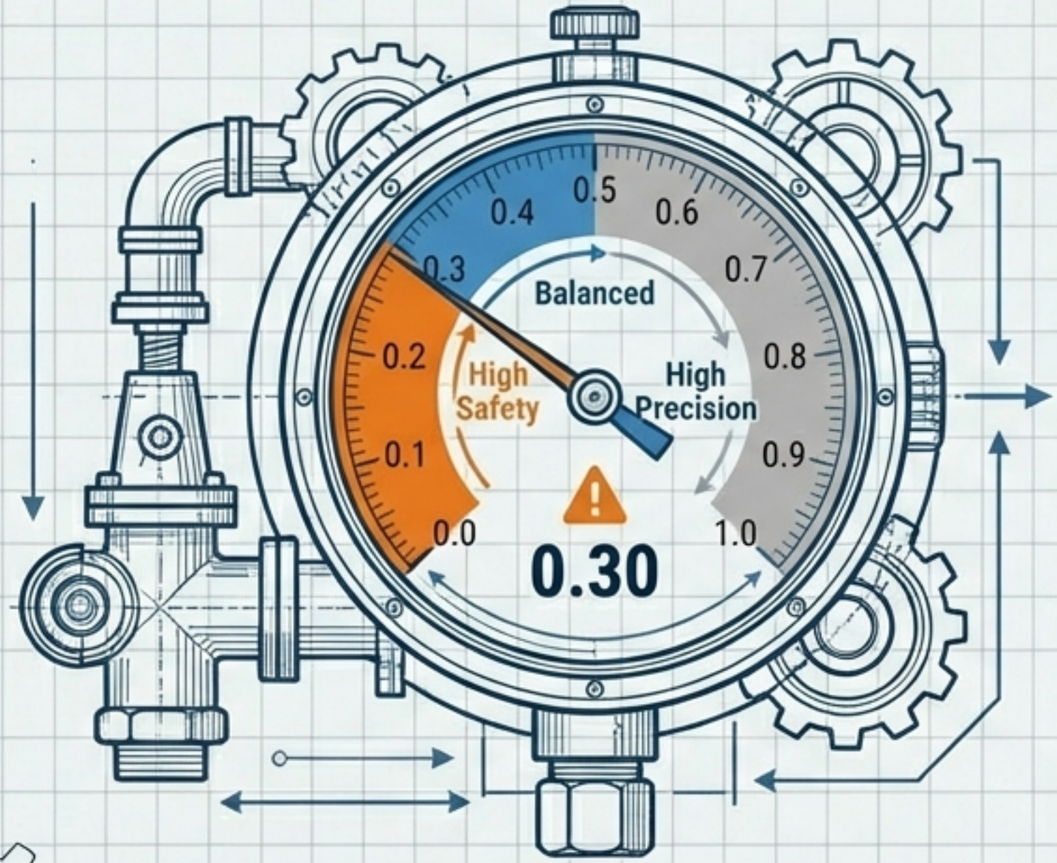
系統優化：調校與負載平衡



Optimization Strategy = Tuning + SMOTE = High Reliability

Noto Sans TC　　Roboto Mono

安全閥調校：決策閾值優化

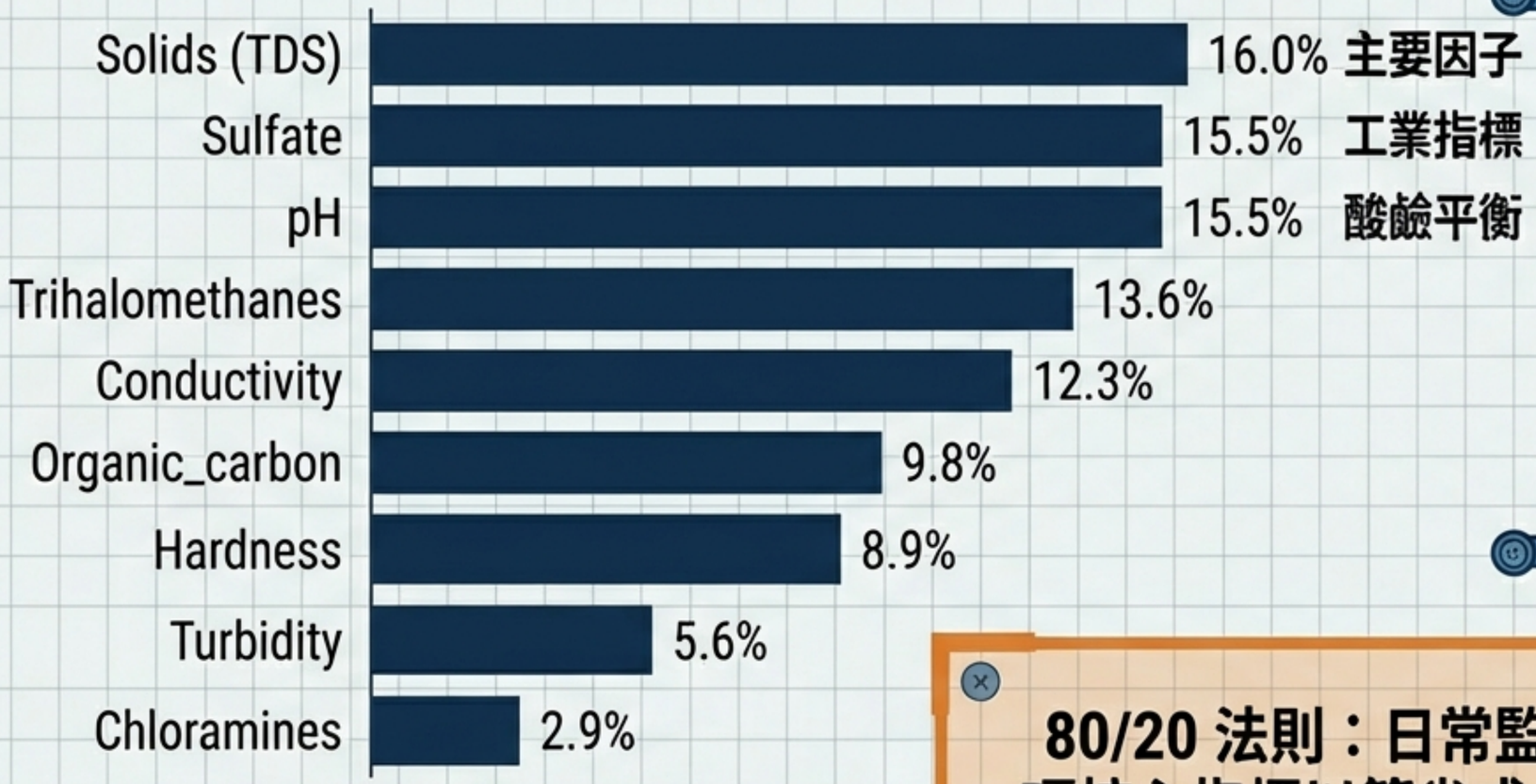


閥值設定後果
(Threshold Consequences)

設定 A (標準)： Threshold **0.50**	Recall: **51%** Precision: **53%**	**漏檢：125 (危險)**
設定 B (平衡)： Threshold **0.42**	Recall: **72%** Precision: **60%**	Youden Index **Best**
設定 C (高安全): Threshold **0.30**	Recall: **91.8%** Precision: **51.3%**	**漏檢：極低 (安全優先)**

建議供水系統採用設定 C (0.30) 以確保公共衛生。

系統驅動因子：特徵重要性

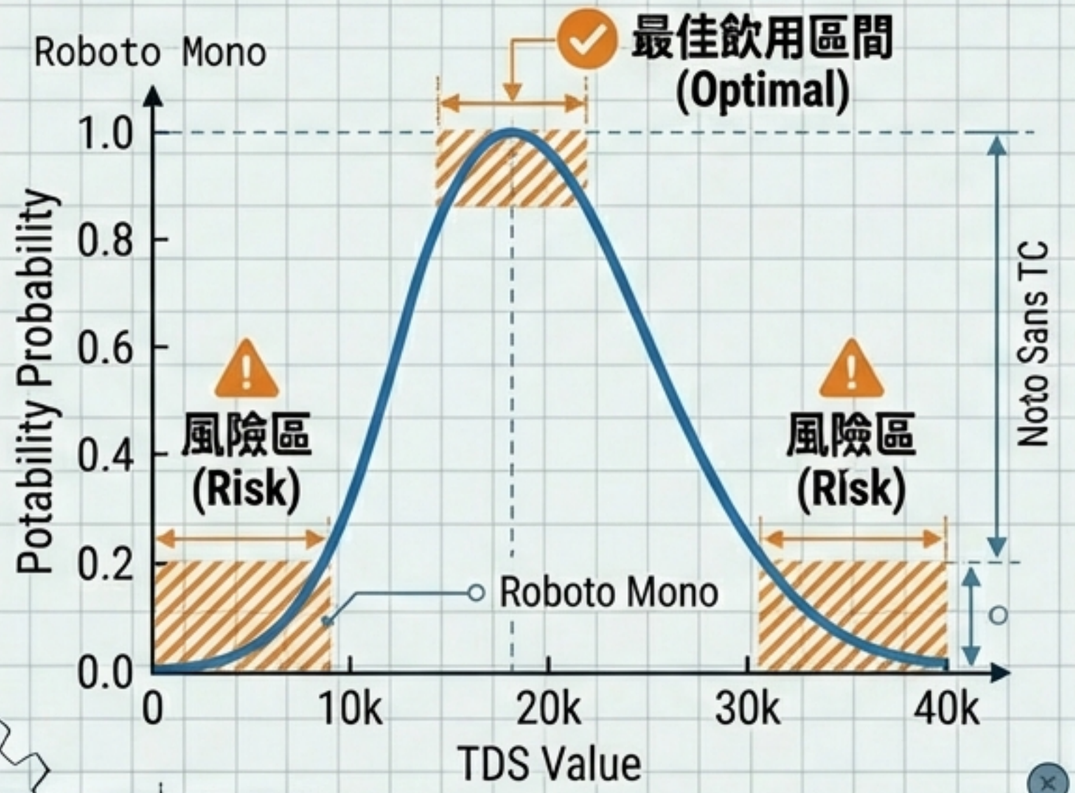


累積貢獻 73%
(Cumulative Power)

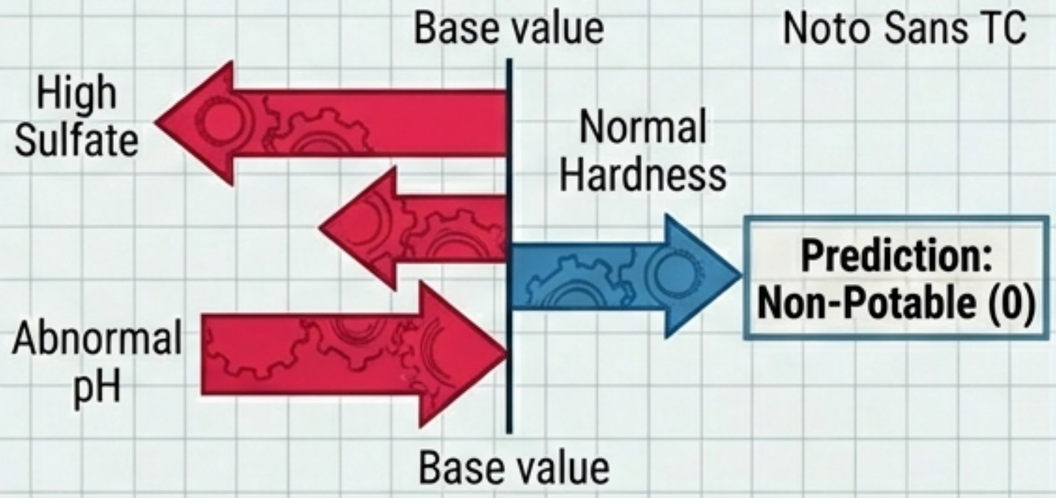
80/20 法則：日常監控可聚焦於前 5 項核心指標以節省成本。

運作邏輯解析：可解釋性 AI (XAI)

Partial Dependence Plot (PDP)

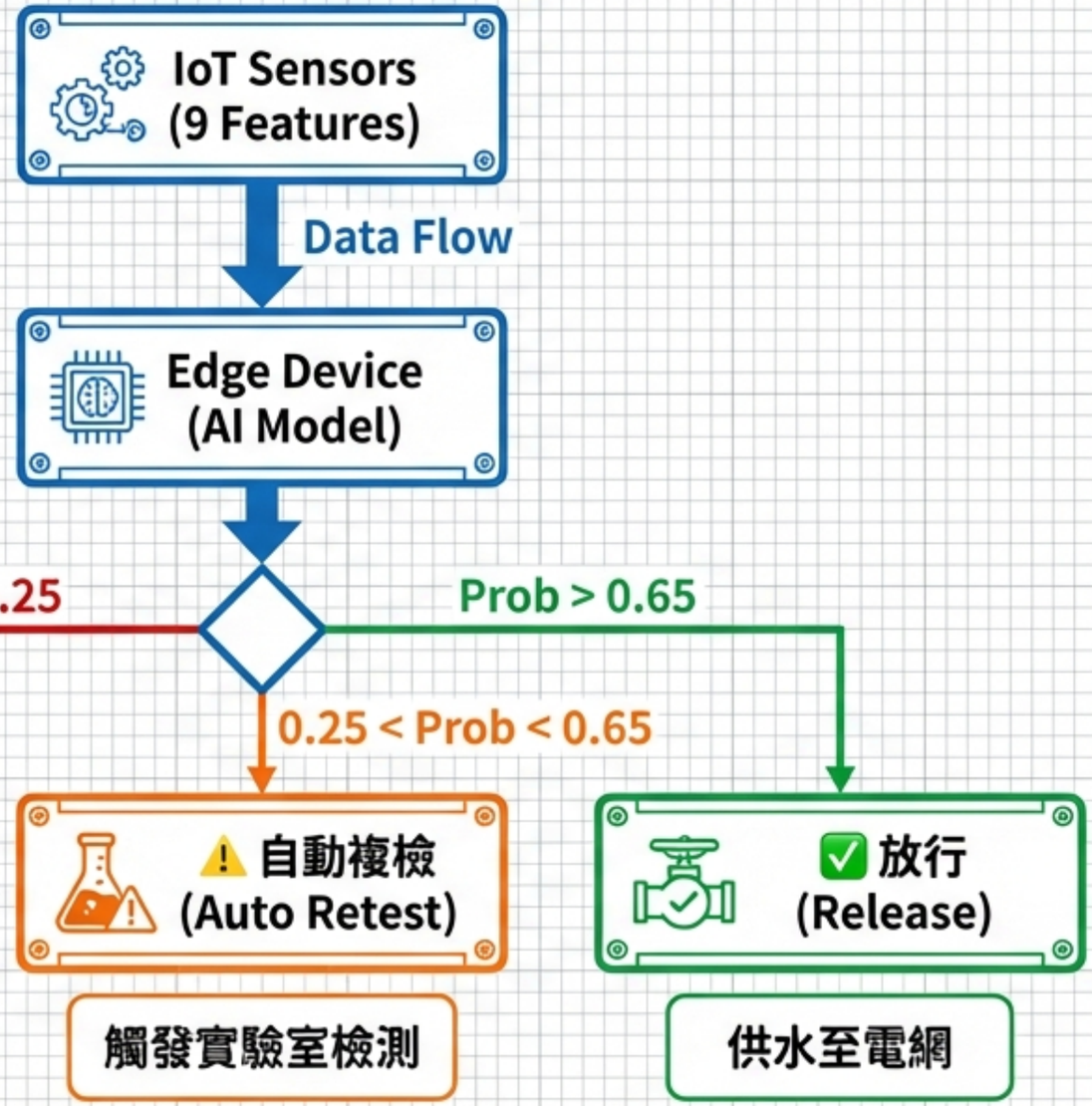


SHAP Force Plot (Sample Breakdown)



模型依據化學原理做出判斷 (Domain Knowledge Validated)。

實務部署架構：分層防禦策略



效益分析：成本與風險平衡

Balance Sheet

傳統實驗室 (Traditional Lab)

成本：NT$ 9,465,000 / 年
人力：高 (High Labor)
速度：48 小時 (Slow)

AI 預測系統 (AI System)

成本：NT$ 2,266,500 / 首年
維護：NT$ 966k / 次年
速度：< 1 秒 (Real-time)

ROI 指標：

節省 (Savings)：76% (Year 1)

回收期 (Payback)：< 3 個月

無形價值：即時攔截污染事件 (Priceless Safety)

專案總結與下一步

未來挑戰 (Future Challenges)：

1. 提升召回率至 > 99%
2. 整合微生物感測器 (E. coli)
3. 邊緣裝置部署 (Edge Deployment)

Full Code available in: **Unit12_Example_WaterQuality.ipynb**