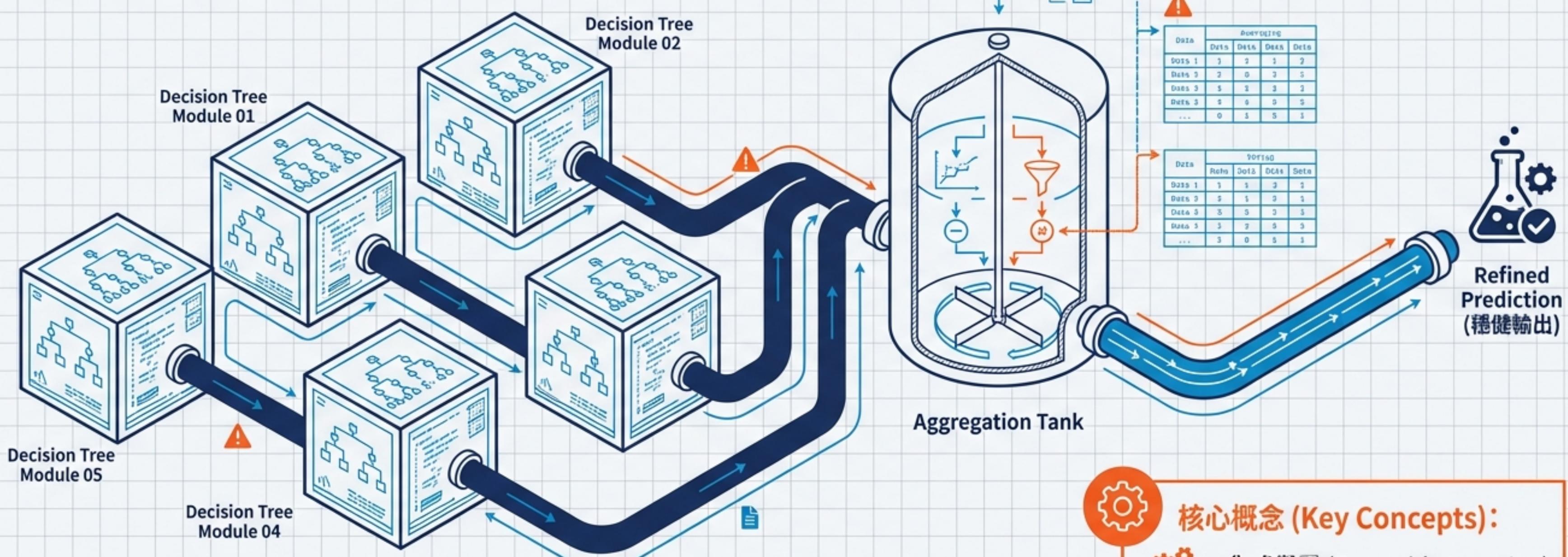


# Unit 13: 隨機森林 (Random Forest)

## 建構穩健的化工預測系統



課程名稱: AI 在化工上之應用 (CHE-AI-114)

授課教師: 莊曜禎 助理教授

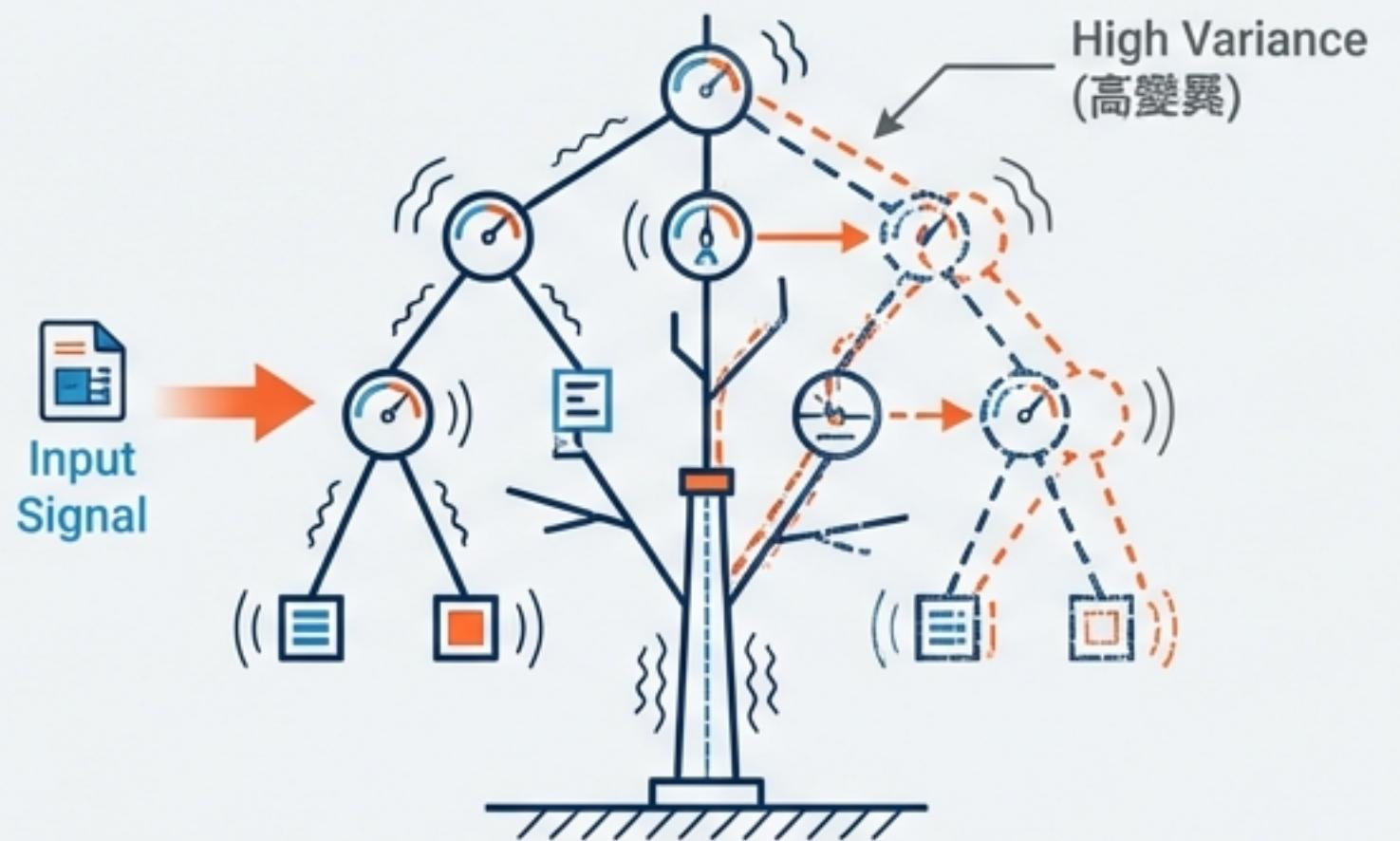


### 核心概念 (Key Concepts):

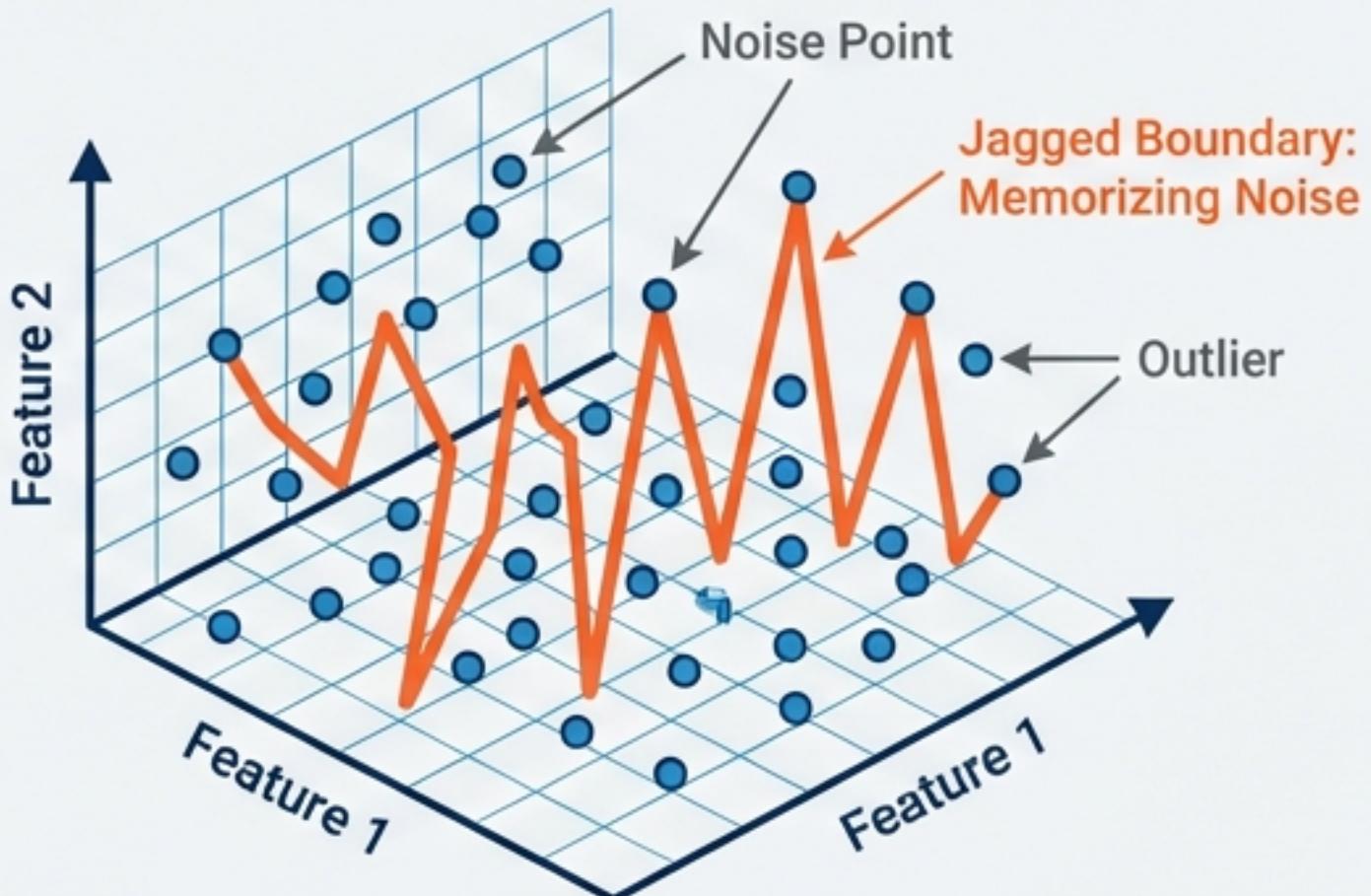
- 集成學習 (Ensemble Learning)
- Bagging 機制
- 穩健性設計

# 單一單元操作的侷限：決策樹的不穩定性

## 結構脆弱性 (Structural Fragility)



## 過擬合現象 (Overfitting)

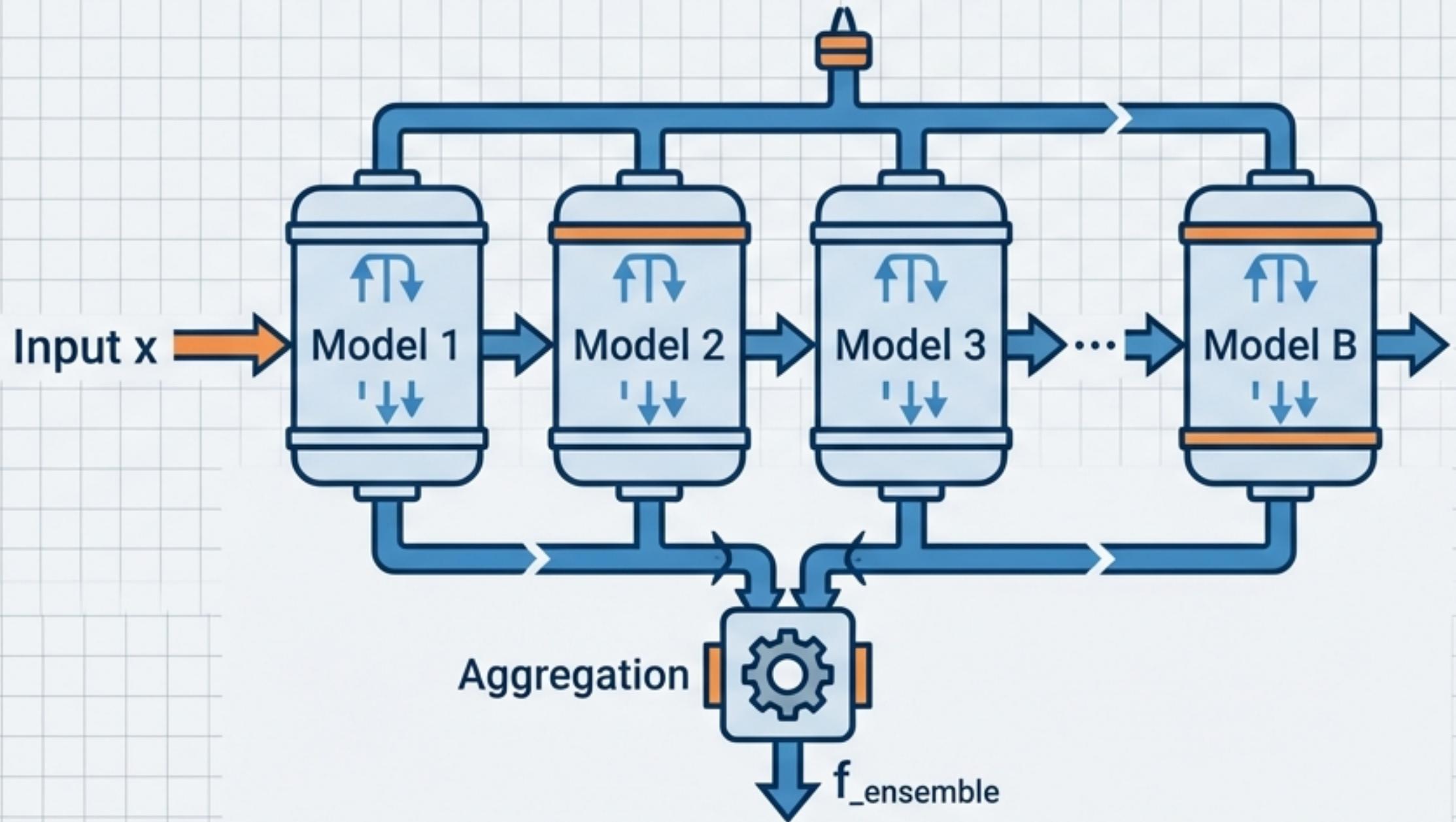


### 問題分析：

1. 高變異性 (High Variance)：訓練資料微小變化導致樹結構大幅改變（如同敏感的感測器）。
2. 容易過擬合 (Overfitting)：無限制生長會記憶訓練集噪音，泛化能力差。
3. 預測不穩定：對新資料的適應力有限。

**Insight:** 就像化工廠不能依賴單一且不穩定的感測器，我們需要一個『冗餘系統』來消除誤差。

# 系統解決方案：集成學習 (Ensemble Learning)



輸出邏輯：

- 回歸 (Regression) → 平均值 (Averaging)
- 分類 (Classification) → 多數決 (Voting)

## 理論基礎

集成公式：

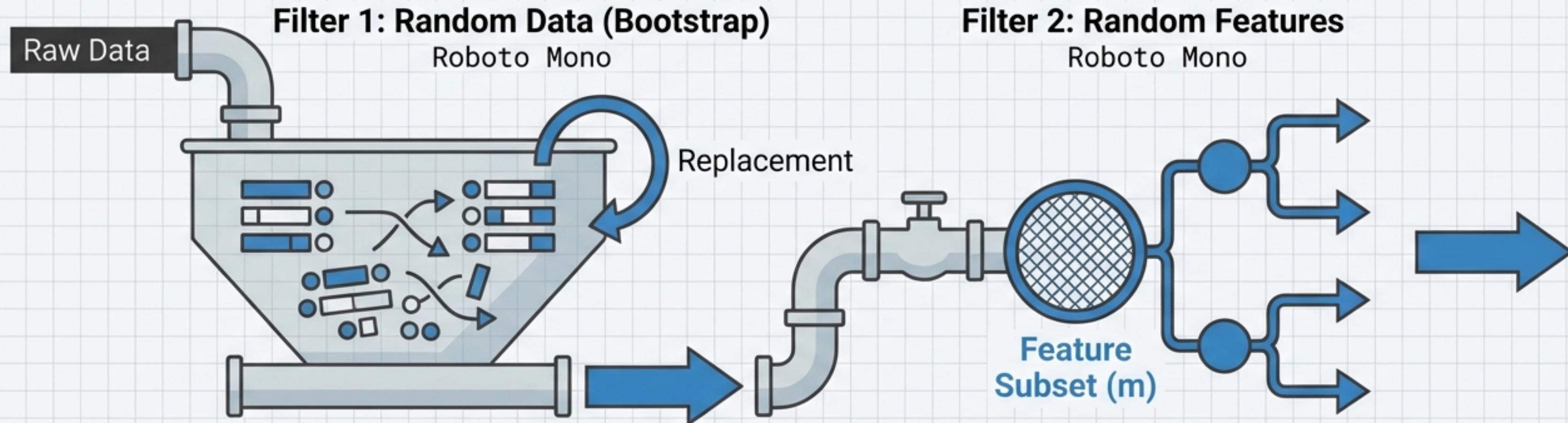
$$f_{\text{ensemble}}(\mathbf{x}) = \frac{1}{B} \sum_b^B f_b(\mathbf{x})$$

變異縮減：

$$\text{Var}_{\text{ensemble}} = \rho \sigma^2 + \left( \frac{1 - \rho}{B} \right) \sigma^2$$

核心機制：Bagging (Bootstrap Aggregating) – 結合多個弱學習器形成強學習器。

# 去關聯機制：隨機性帶來的多樣性

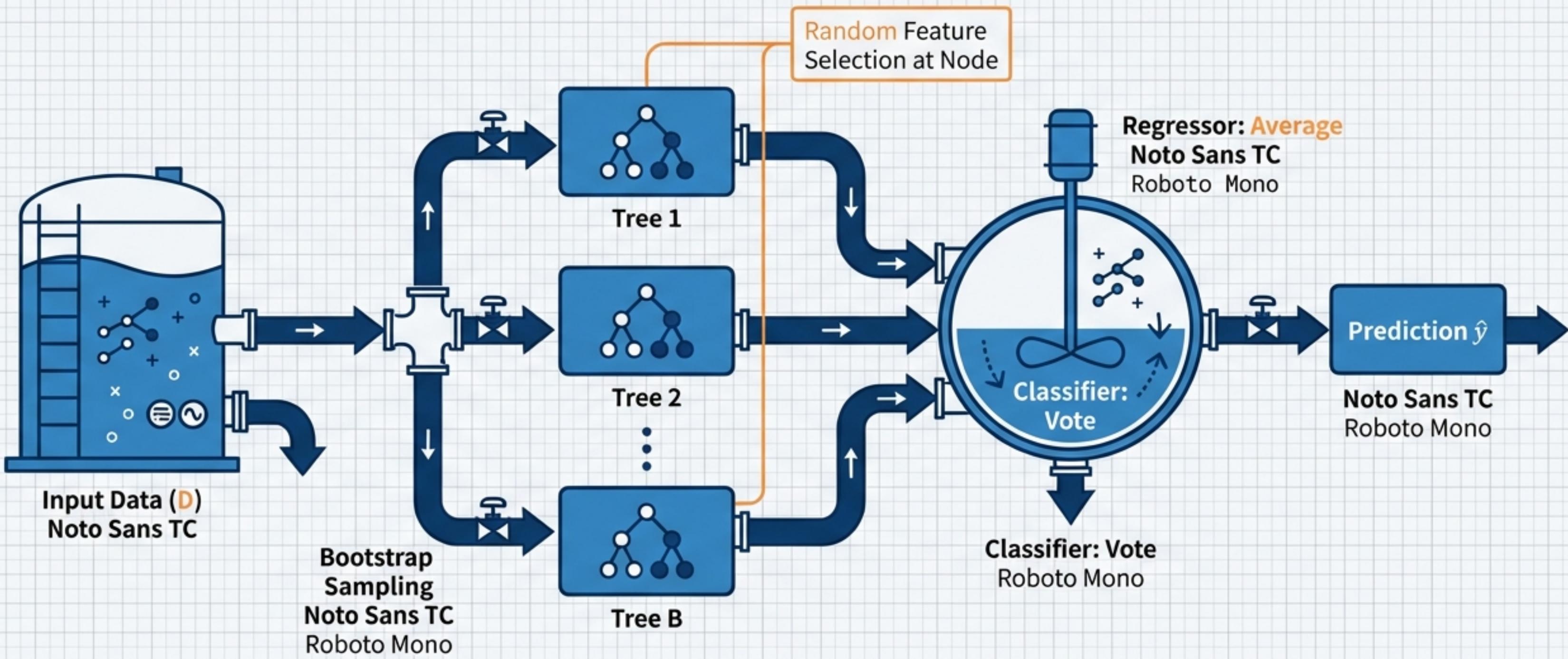


從原始數據有放回地隨機採樣。約 63.2% 樣本被選中，剩餘 36.8% 為 Out-of-Bag (OOB) 樣本。

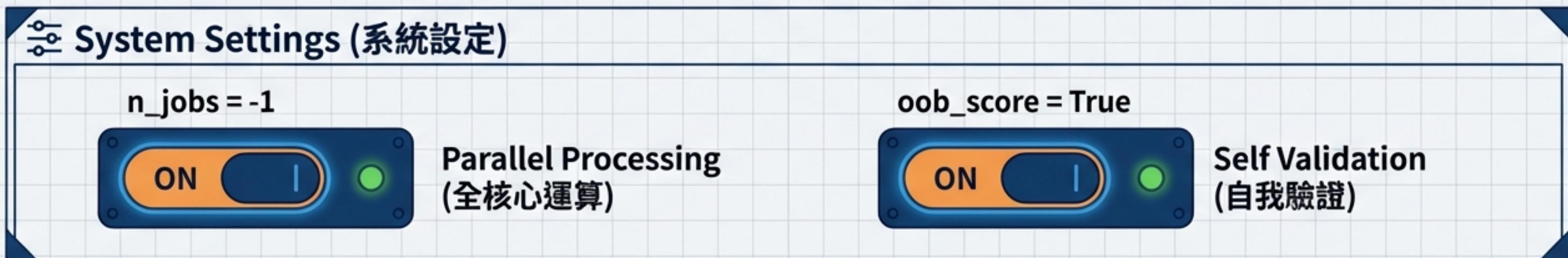
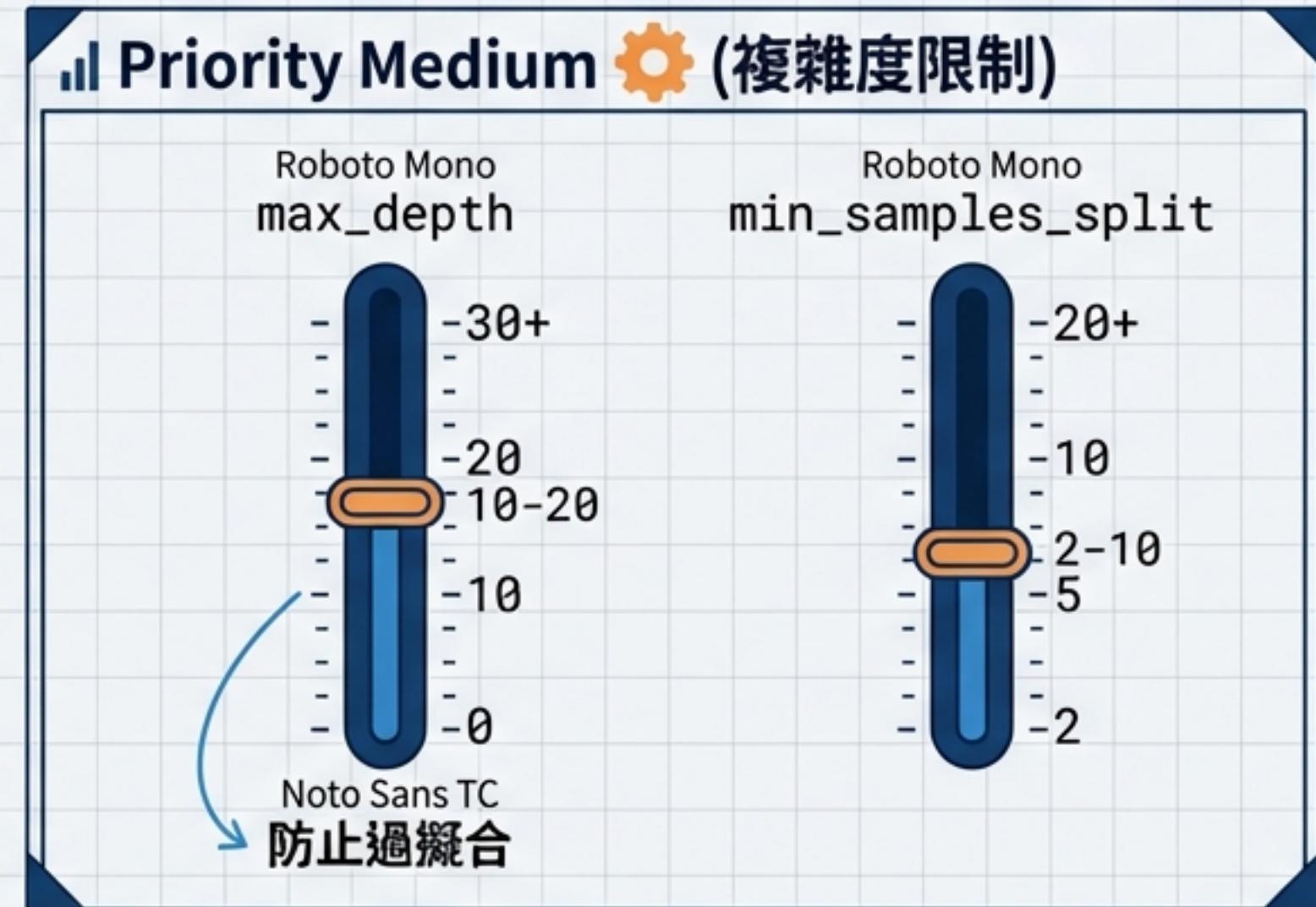
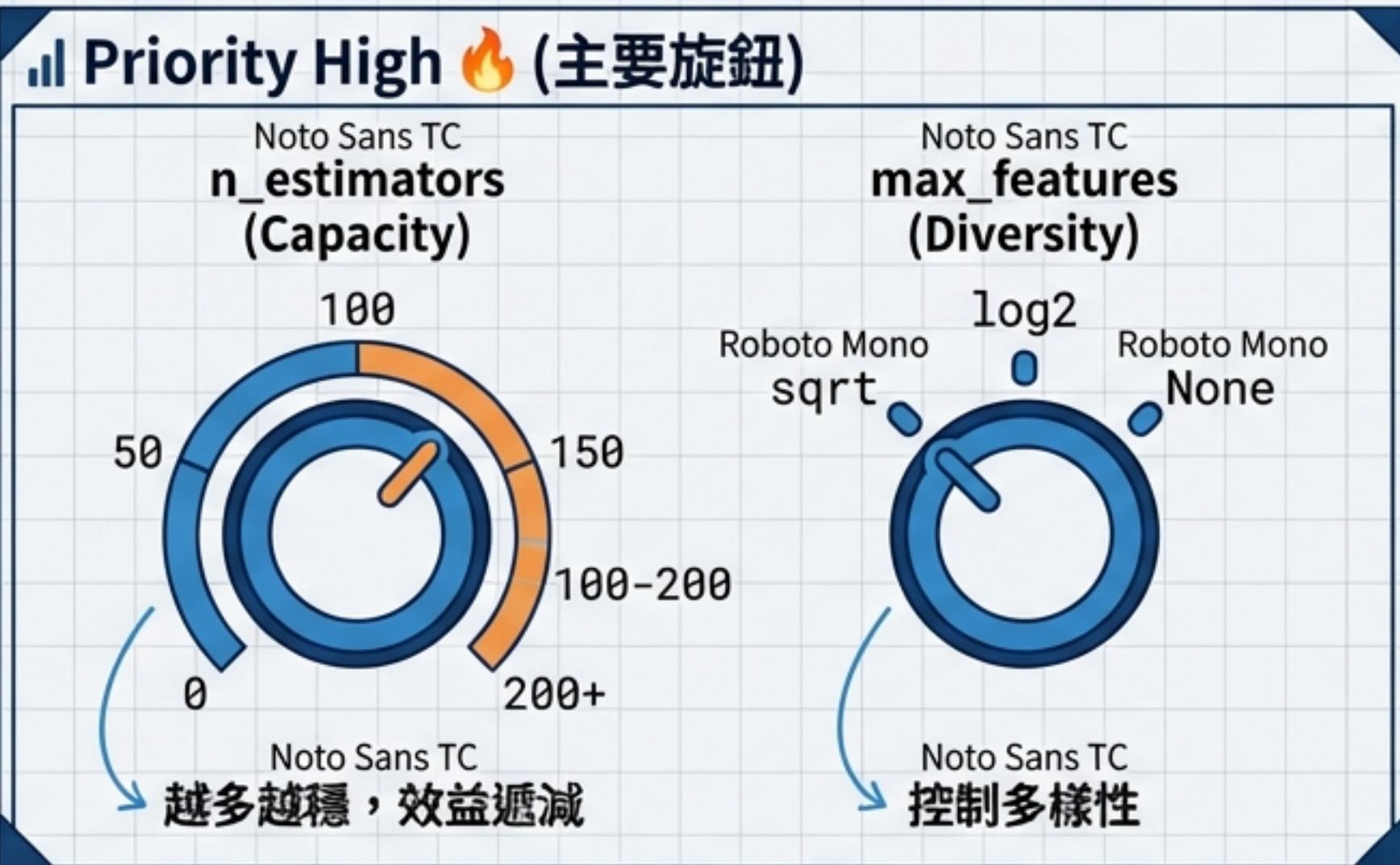
每次節點分裂時，僅考慮隨機子集特徵  $m$ 。避免單一強特徵主導。  
回歸  $m = p/3$ ，分類  $m = \sqrt{p}$

**Engineering Insight** | Decorrelation (去關聯) 是關鍵：若所有反應器都使用相同的進料和設定，誤差會疊加。隨機性迫使模型從不同角度觀察數據，實現『 $1+1 > 2$ 』的效果。

# 完整算法流程 (Process Flow Diagram)



# 控制面板：超參數調整 (Hyperparameters)



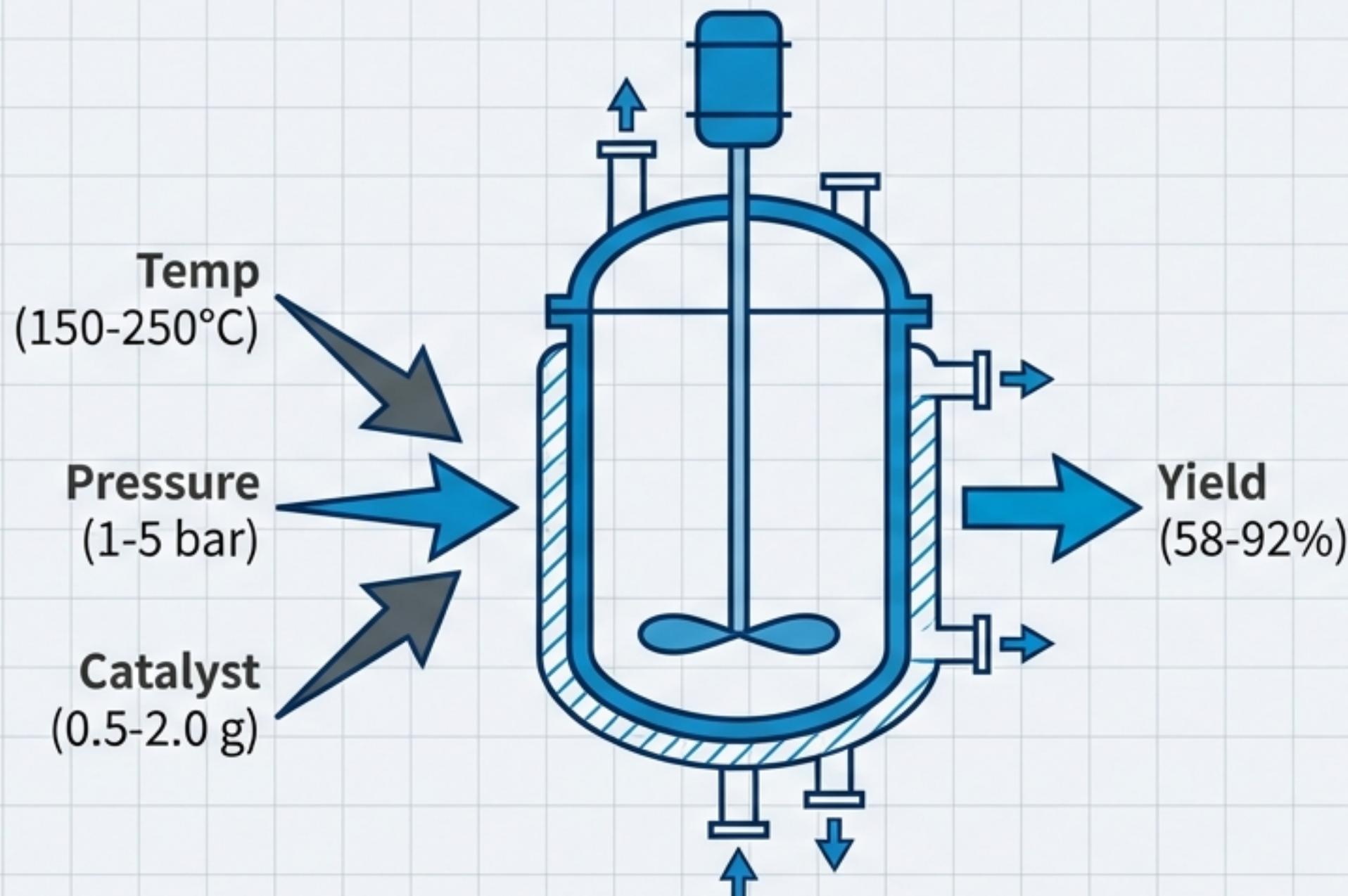
# 系統實作：Scikit-Learn 代碼

```
from sklearn.ensemble import RandomForestRegressor

# 1. 設定參數 (Configuration)
rf_model = RandomForestRegressor(
    n_estimators=100,                  # 100 棵樹
    max_depth=15,                     # 限制深度
    max_features='sqrt',              # 隨機特徵
    oob_score=True,                   # 開啟 OOB 驗證
    random_state=42,
    n_jobs=-1                         # 並行加速
)
# 2. 啟動系統 (Train)
rf_model.fit(X_train, y_train)
# 3. 讀取狀態 (Validate)
print(f"OOB Score: {rf_model.oob_score_:.4f}")
```

Syntax Note: 與標準 Decision Trees 語法完全一致，升級無痛。

# 實地測試 I：反應器產率優化 (Regression)



## Test Report

Data Scale: 1000 實驗樣本 (Non-linear)

Model Performance:

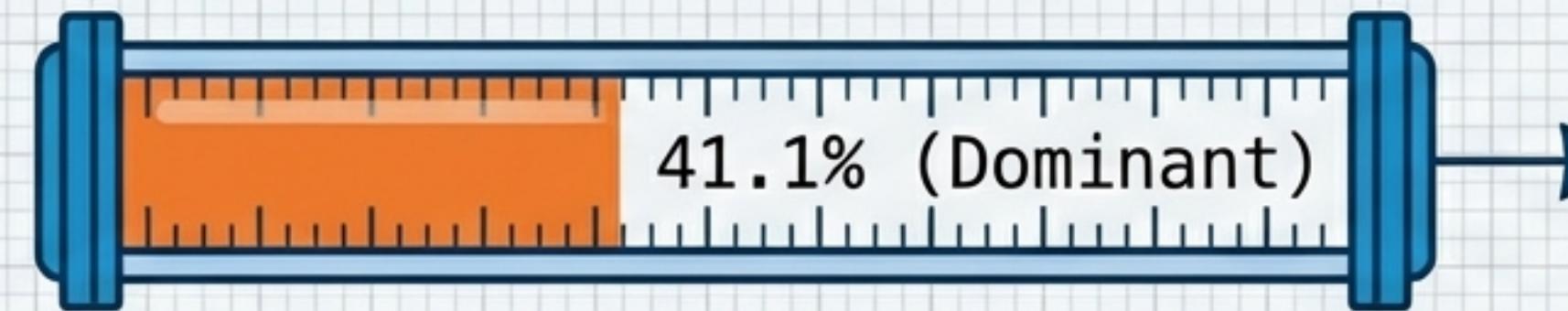
- Test R<sup>2</sup> Score: 0.73
- RMSE (Error): 3.48%

Conclusion:

Random Forest outperforms linear models by capturing the complex 'sweet spot' of temperature and pressure.

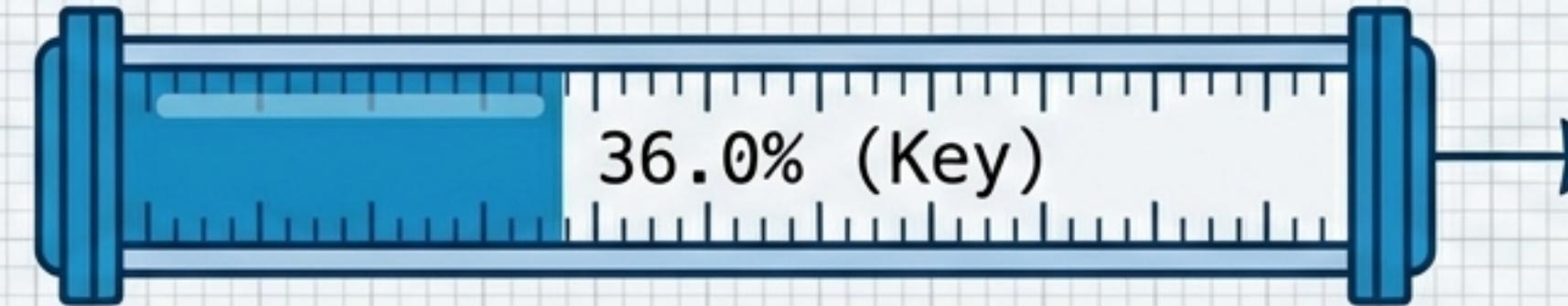
# 關鍵變數識別：特徵重要性 (Feature Importance)

Pressure



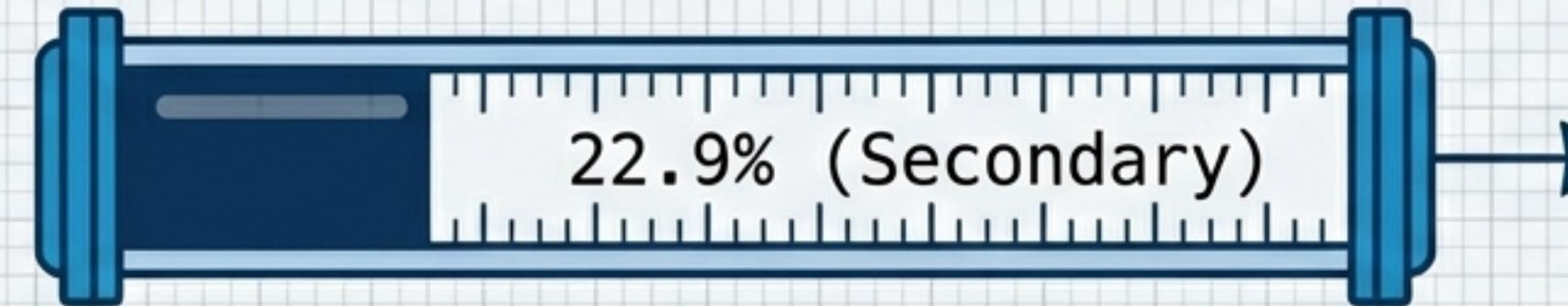
直接影響氣相反應速率，  
變異性最大。

Temperature



具有最適區間（非線性），  
過高過低皆不利。

Catalyst

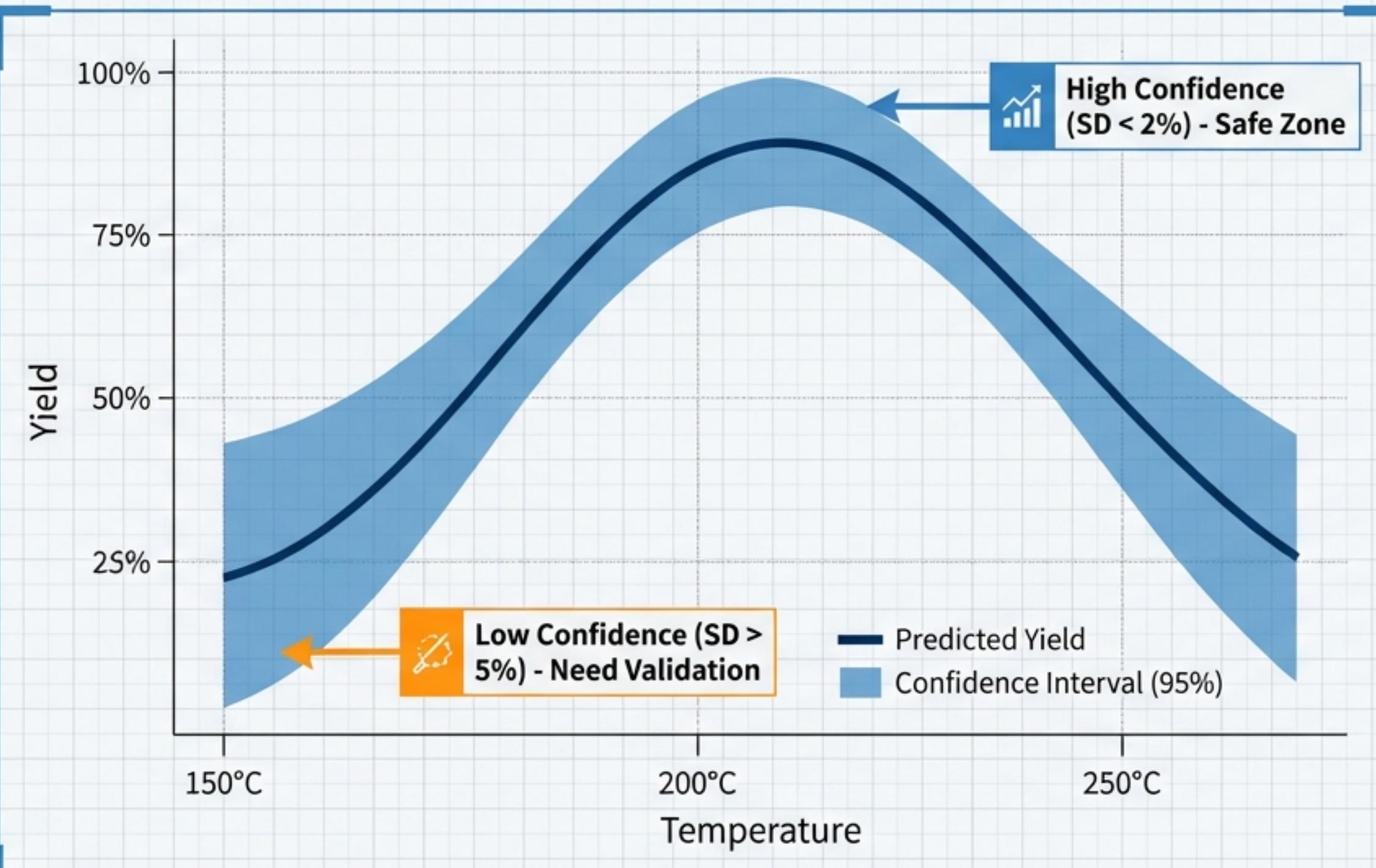


在實驗範圍內接近飽和，  
邊際效益較低。



Takeaway: 模型驗證了物理直覺 (Physics-informed ML)。

# 風險評估：不確定性量化 (Uncertainty Quantification)



## 機制 (Mechanism):

利用各樞紐預測值的標準差 (SD) 估計信心區間。

Formula:

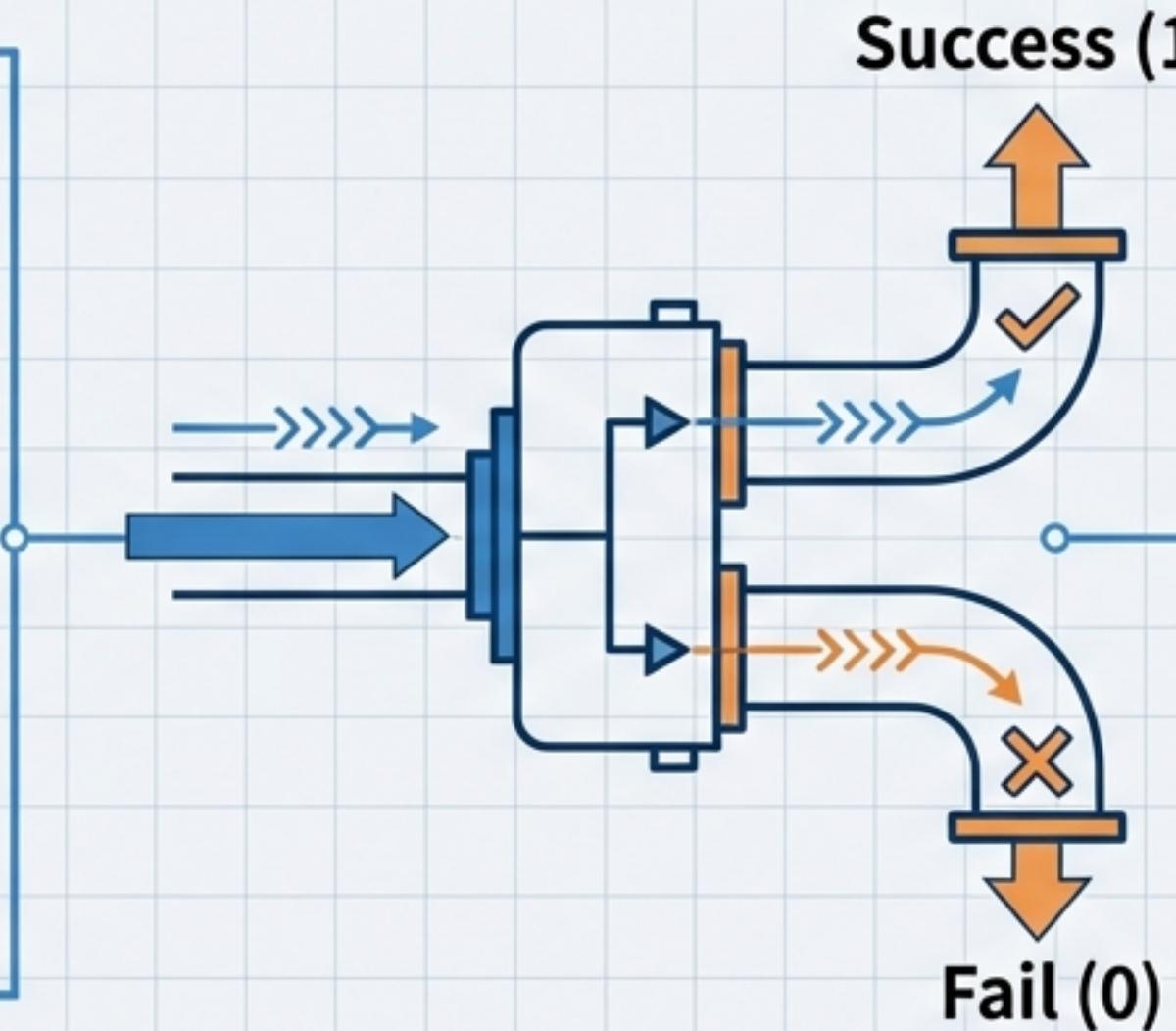
$$\text{Interval} = \hat{y} \pm 1.96 \times \text{SD}$$



# 實地測試 II：反應成功預測 (Classification)

**Scenario:**

- **Goal:** Binary Classification (Success/Fail)
- **Features:** Temp, Pressure, Catalyst, Ratio, Time
- **Data Balance:** 56% Fail / 44% Success



**Performance Metrics:**

- **Accuracy:** 67.75% (優於隨機猜測)
- **AUC Score:** 0.72 (中等區分能力)
- **Stability:** Cross-Validation Std < 1%



**Key Insight:** GridSearch 調參後性能提升，且模型展現極高穩定性。

# 性能規格比較：RF System vs. Single Unit

項目 (Item)	單棵決策樹 (Single Tree)	隨機森林 (Random Forest)
R <sup>2</sup> Score	0.62	0.73 (Winner)
RMSE (誤差)	4.1%	3.5% (Winner)
穩定性	差 (High Variance)	優 (Ensemble Avg)
過擬合風險	高	低 (Bagging)
訓練時間	極快 (0.05s)	較慢 (0.30s)
可解釋性	直觀可視化	特徵重要性排序

 隨機森林以些微的計算成本換取了顯著的準確度與穩定性提升。

# 工程評估：優勢與操作限制



## 優勢 (Pros)

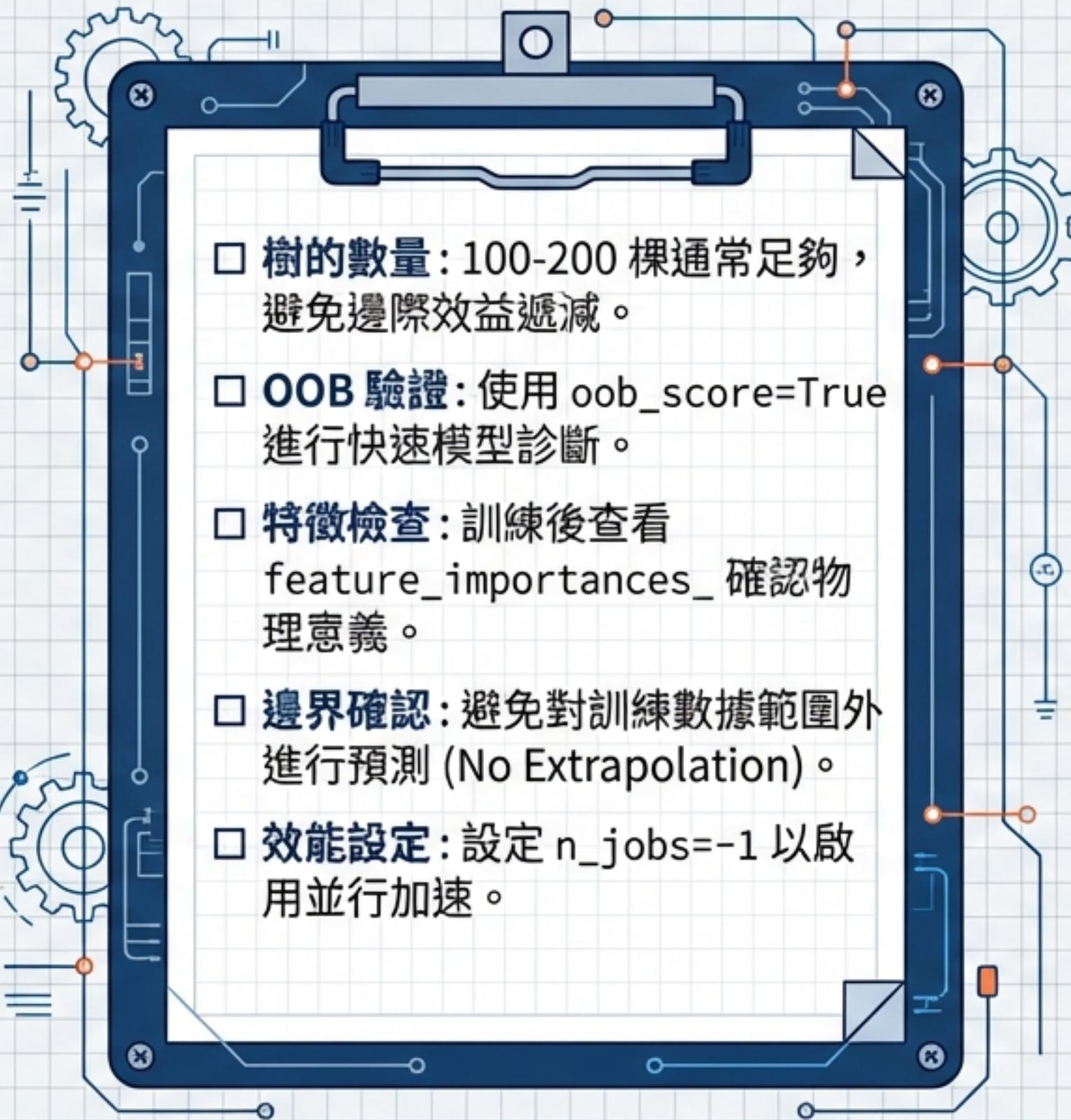
- ✓ 1. **Robustness** : 對異常值 (Outliers) 和噪聲不敏感。
- ✓ 2. **Non-linearity** : 自動捕捉複雜交互作用，無需特徵縮放。
- ✓ 3. **Validation** : OOB Score 可作為免費的 Cross-Validation。



## 限制 (Cons)

- ⚠ 1. **Black Box** : 內部結構難以完全透視。
- ⚠ 2. **No Extrapolation** : 無法預測訓練範圍以外的數值 (如預測未見過的高產率)。
- ⚠ 3. **Cost** : 樹越多，預測與訓練越慢。

# 實務操作清單 (Best Practices Checklist)



# 總結：從單點到系統的進化



Decision Tree  
(Unit)



Bagging Logic



Random Forest  
(System)

## Key Takeaways

- **1. 原理**: Bagging + 隨機特徵 = 降低變異 (Variance Reduction)。
- **2. 應用**: 適合處理化工製程中的複雜非線性與噪聲數據。
- **3. 下一步**: 開啟 `Unit13\_Random\_Forest\_Lab.ipynb` 進行實作。