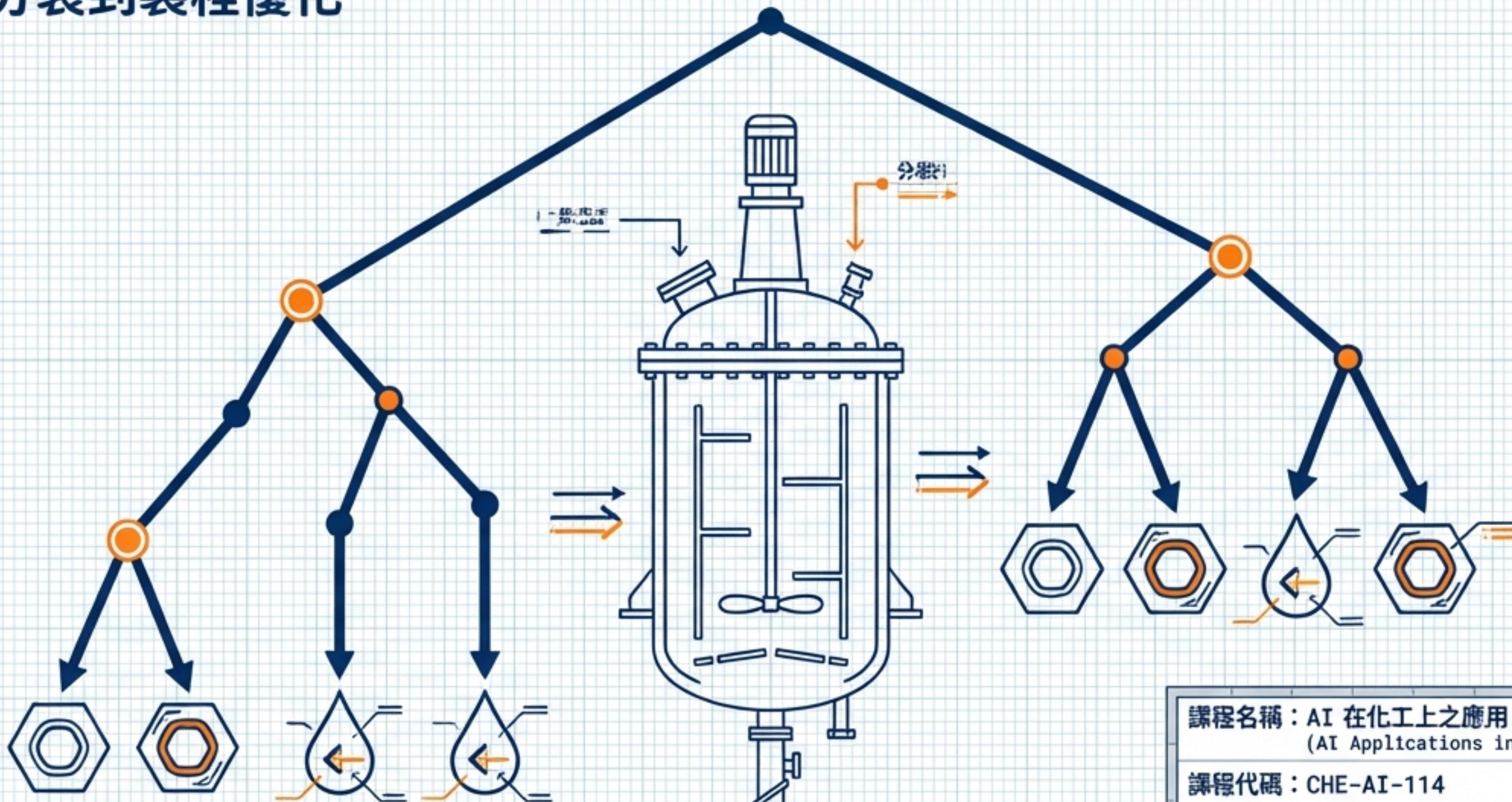


Unit 11: 決策樹回歸 (Decision Tree Regression)

從數據分裂到製程優化



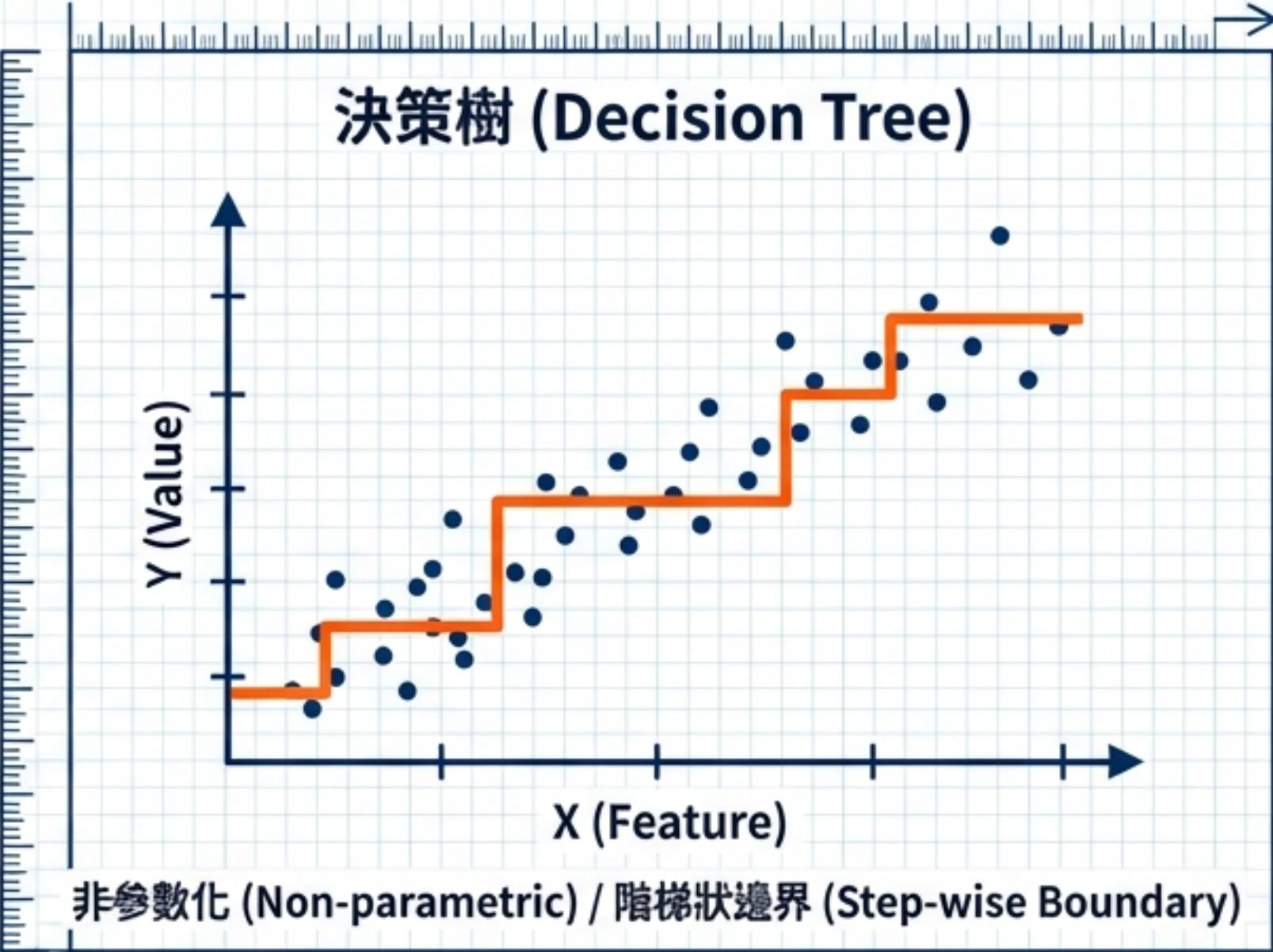
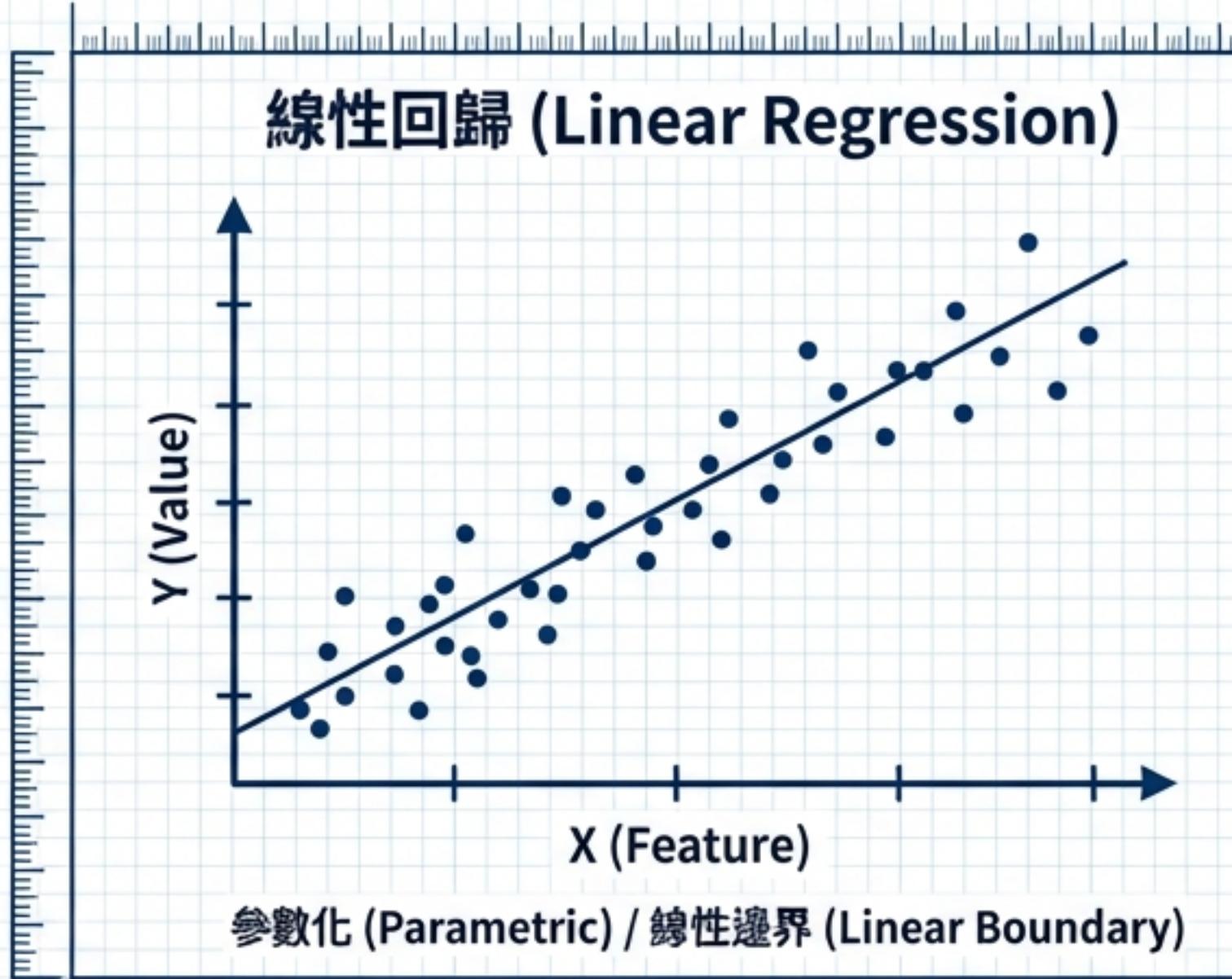
課程名稱：AI 在化工上之應用
(AI Applications in Chemical Engineering)

課程代碼：CHE-AI-114

授課教師：莊曜楨 助理教授

所屬單位：逢甲大學 化工系 智慧程序系統工程實驗室

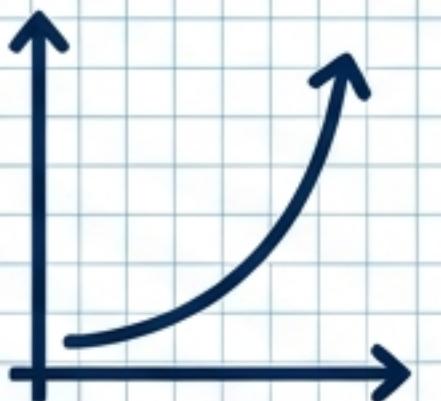
什麼是決策樹回歸？(What is Decision Tree Regression?)



- 一種基於樹狀結構的非參數監督式學習方法。
- 通過一系列的 **二元決策規則 (Binary Decision Rules)** 將特徵空間遞迴地分割。
- 預測值 = 該區域內所有樣本的 **平均值 (Mean)**。

不是公式，而是流程圖

為何化工需要決策樹？(Why Use Trees in ChE?)



捕捉非線性 (Non-linearity)

化學反應速率通常是指數關係 (Arrhenius Equation)，線性模型無法擬合。



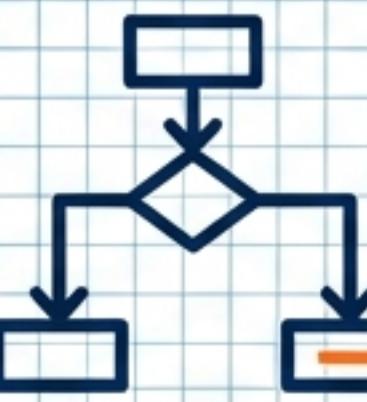
交互作用 (Interactions)

自動發現變數間的協同效應
(例如：只有在高溫下，壓力才影響產率)。



魯棒性 (Robustness)

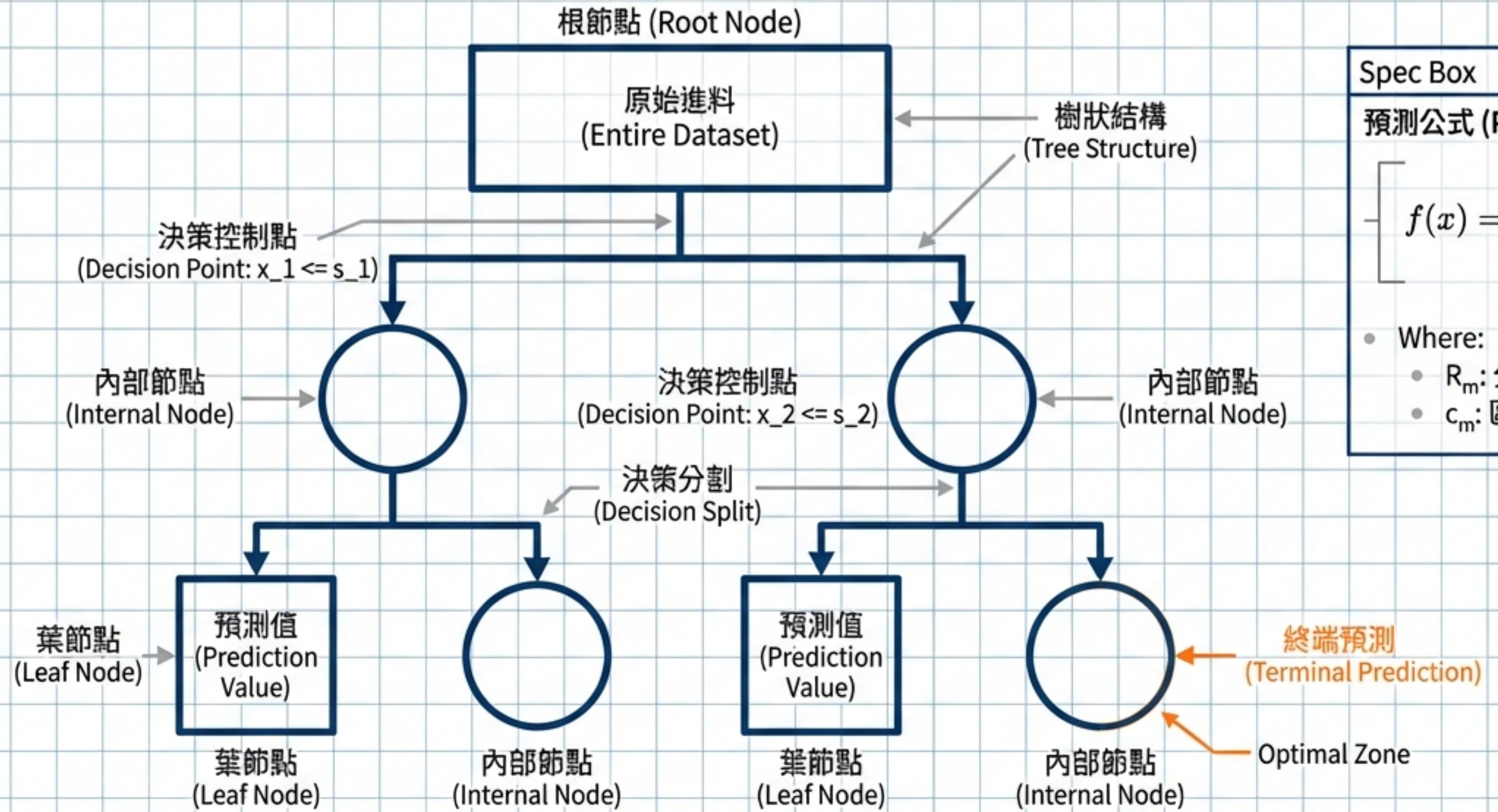
對異常值 (Outliers) 較不敏感，且無需對數據進行標準化 (No Scaling Needed)。



可解釋性 (Interpretability)

產生的規則 ("If T > 200...") 直觀且符合工程邏輯。

樹的解剖學：結構與數學表示 (The Anatomy of a Tree)



Spec Box

預測公式 (Prediction Formula):

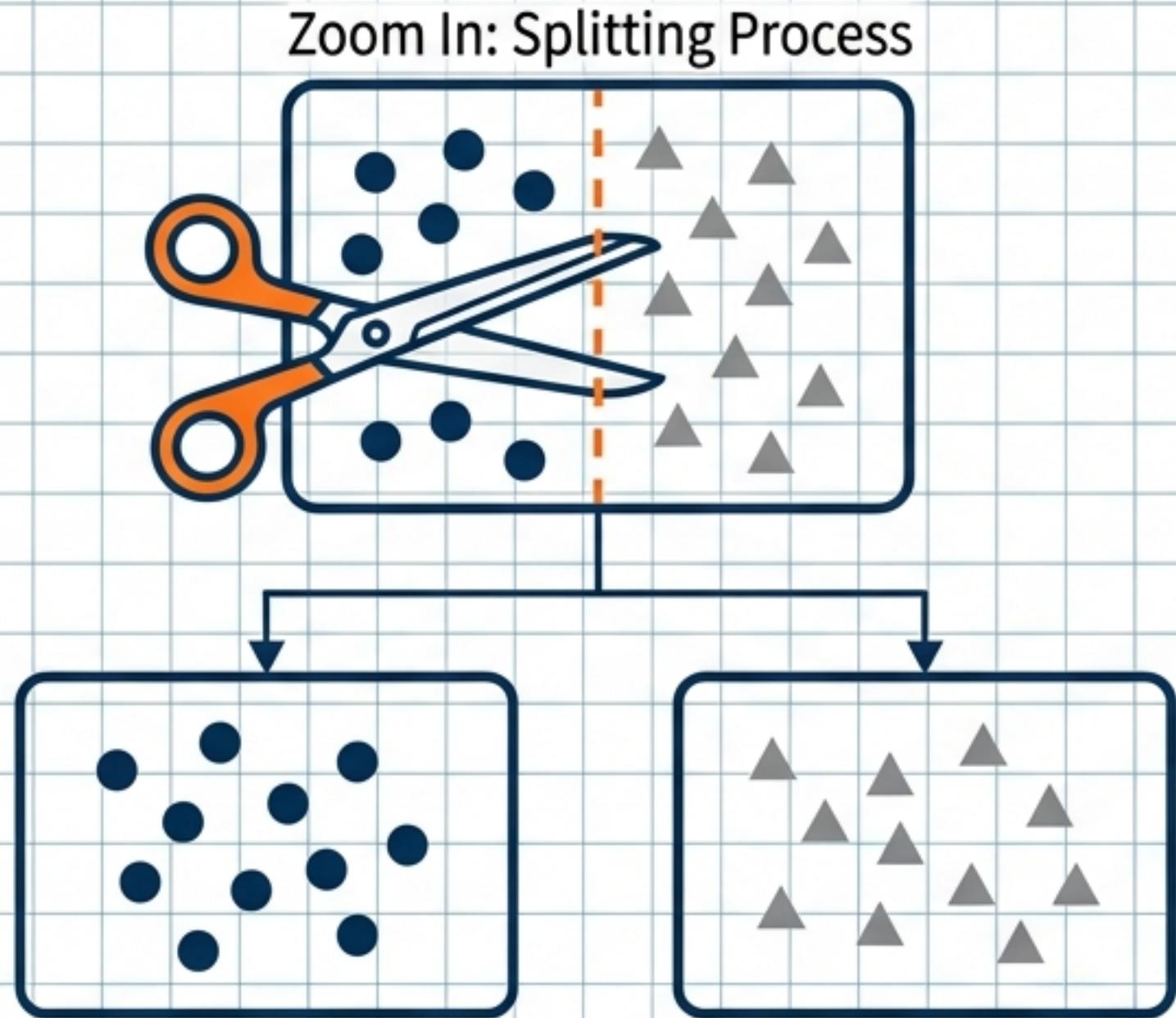
$$f(x) = \sum_m c_m \cdot 1(x \in R_m)$$

Where:

- R_m : 分割後的區域 (Region)
- c_m : 區域內的平均值 (Region Mean)

生長機制：分裂準則 (Splitting Criteria)

CART 演算法與最小化誤差 (CART & Error Minimization)



核心邏輯 (Core Logic)

貪婪演算法 (Greedy Algorithm) – 尋找能使
“雜質” 減少最多的特徵與切點。

目標函數 (Objective Function: MSE)

目標：最小化均方誤差 (Minimize MSE)

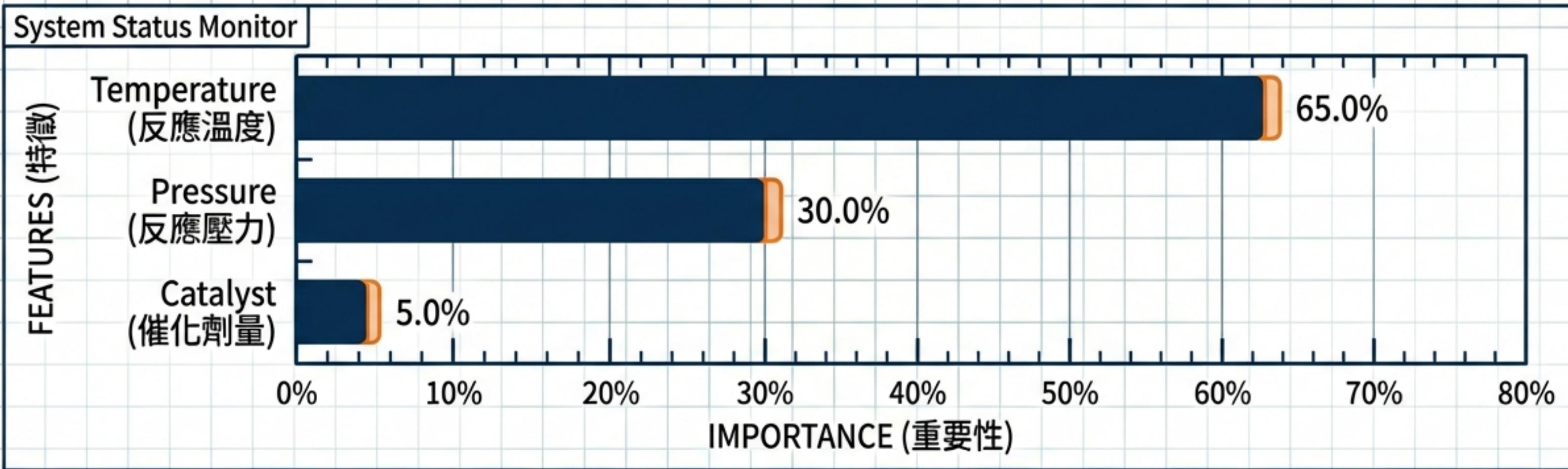
$$(j^*, s^*) = \arg \min [N_{\text{left}} \cdot \text{MSE}_{\text{left}} + N_{\text{right}} \cdot \text{MSE}_{\text{right}}]$$

解釋 (Explanation)

演算法遍歷所有特徵 (j) 與所有閾值 (s)，找出
最佳分割點。

特徵重要性 (Feature Importance)

誰在主導製程？(Who Drives the Process?)

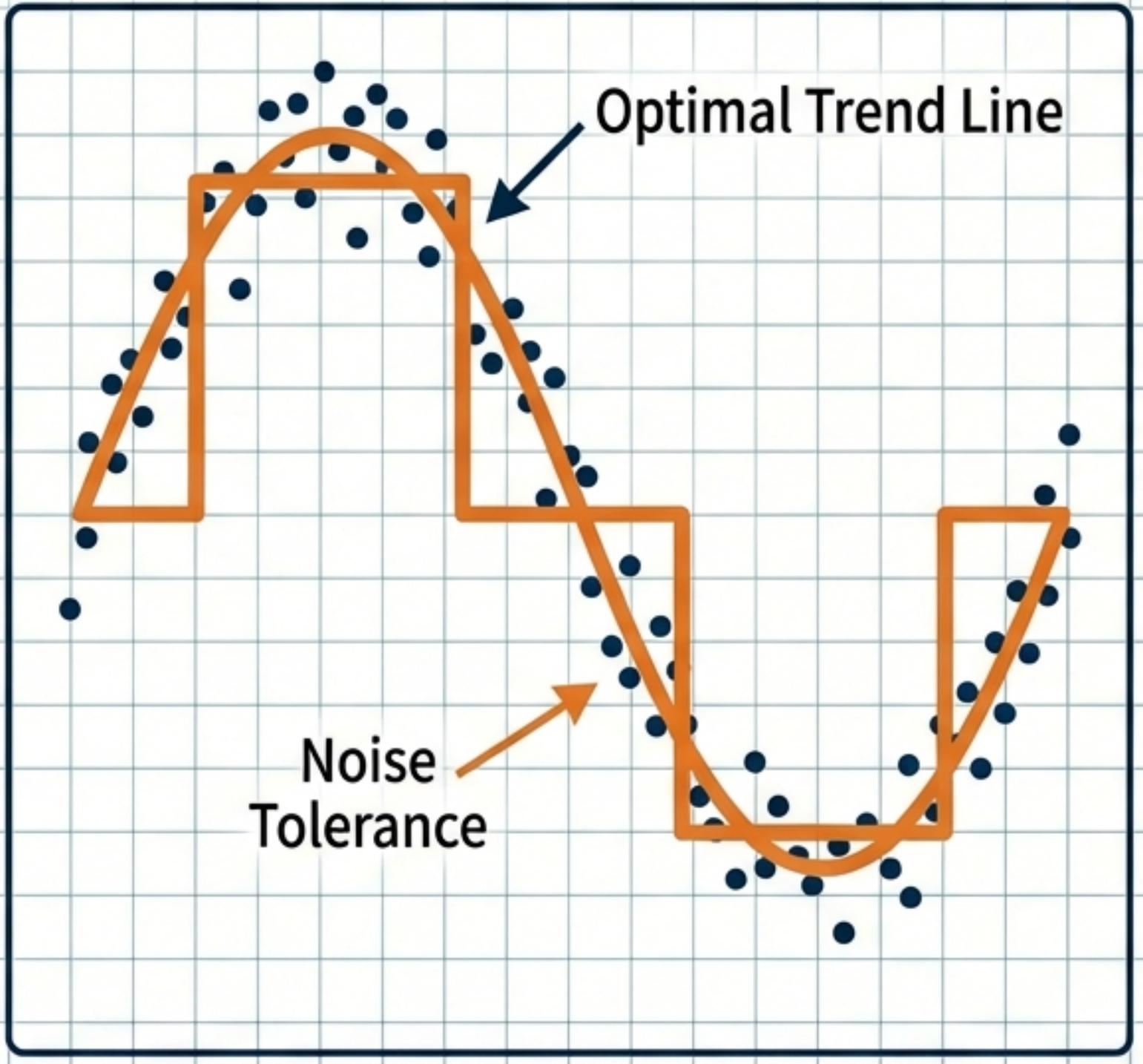


定義：特徵重要性 = 該特徵在所有節點分裂中貢獻的 **總雜質減少量 (Total Impurity Reduction)**。

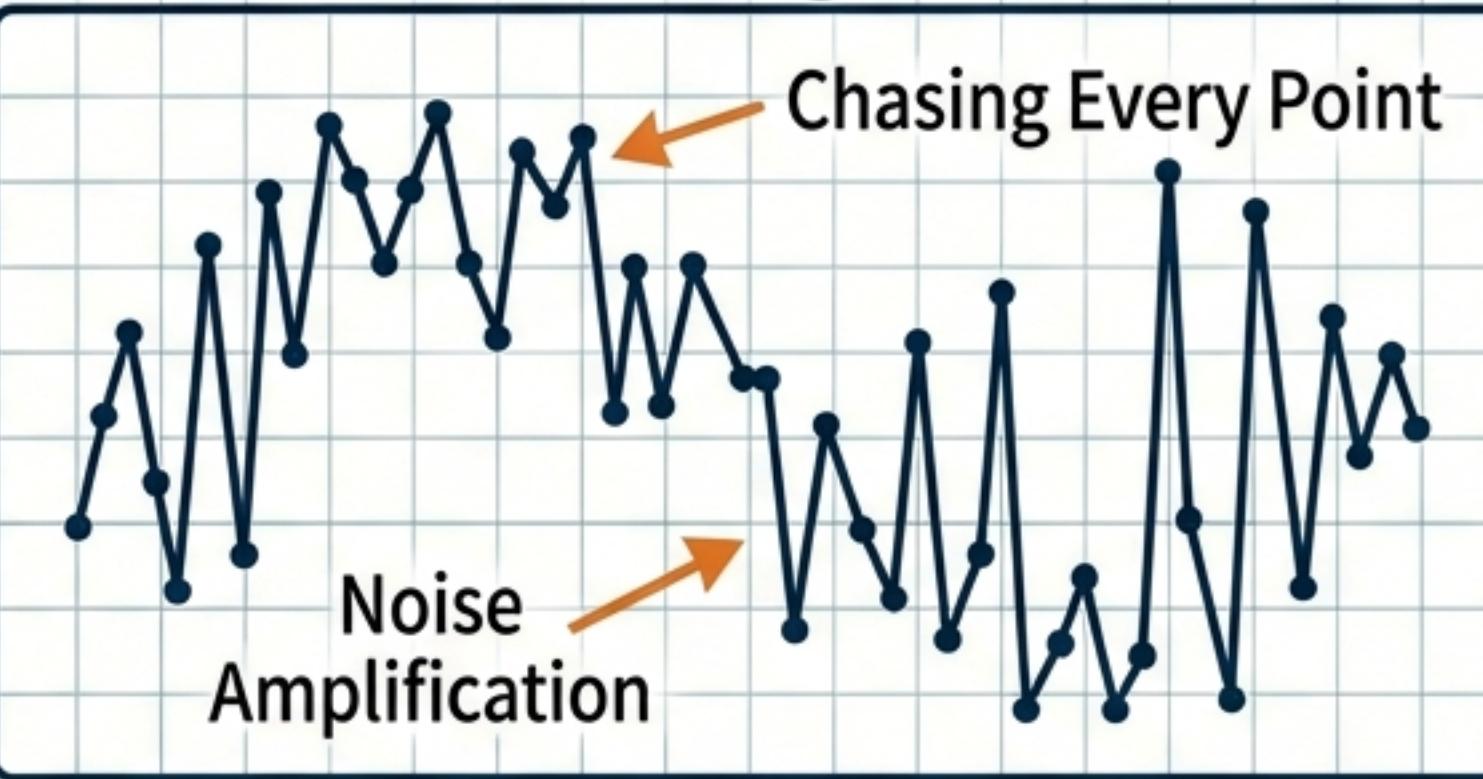
工程洞察 (Engineering Insight)：重要性高的特徵 (High Importance) 代表它是頻繁用於決策樹上層的關鍵變數。這是化工工程師理解‘黑箱’的第一步。

潛在危險：過擬合 (The Enemy: Overfitting)

理想模型 (Good Generalization)



過擬合 (Overfitting)

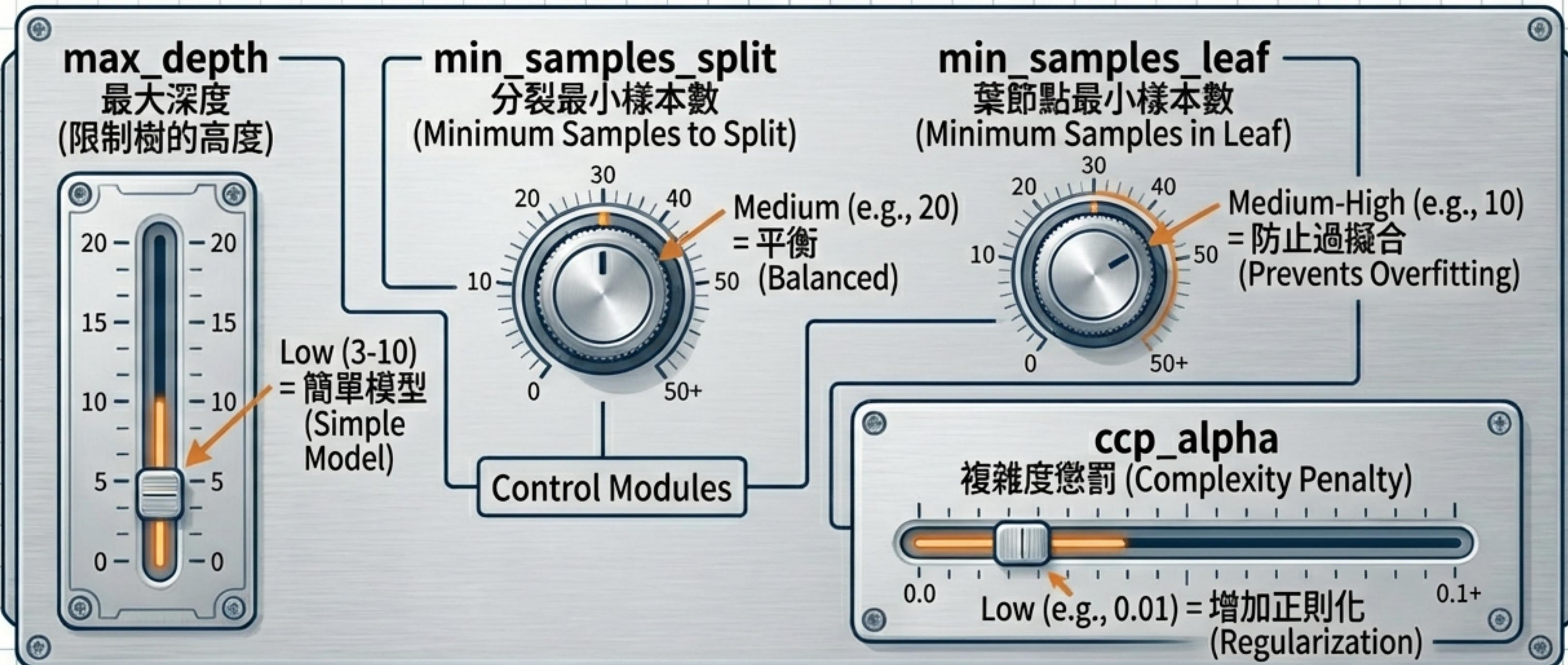


WARNING

症狀 (Symptoms):

- 訓練集 $R^2 \approx 1.0$ (完美擬合) ✓
- 測試集 $R^2 <<$ 訓練集 (實際表現差) ✗
- 類比：就像控制器 (Controller) 過度反應於感測器的噪音，而非製程趨勢。

控制複雜度：超參數調校 (Hyperparameters)

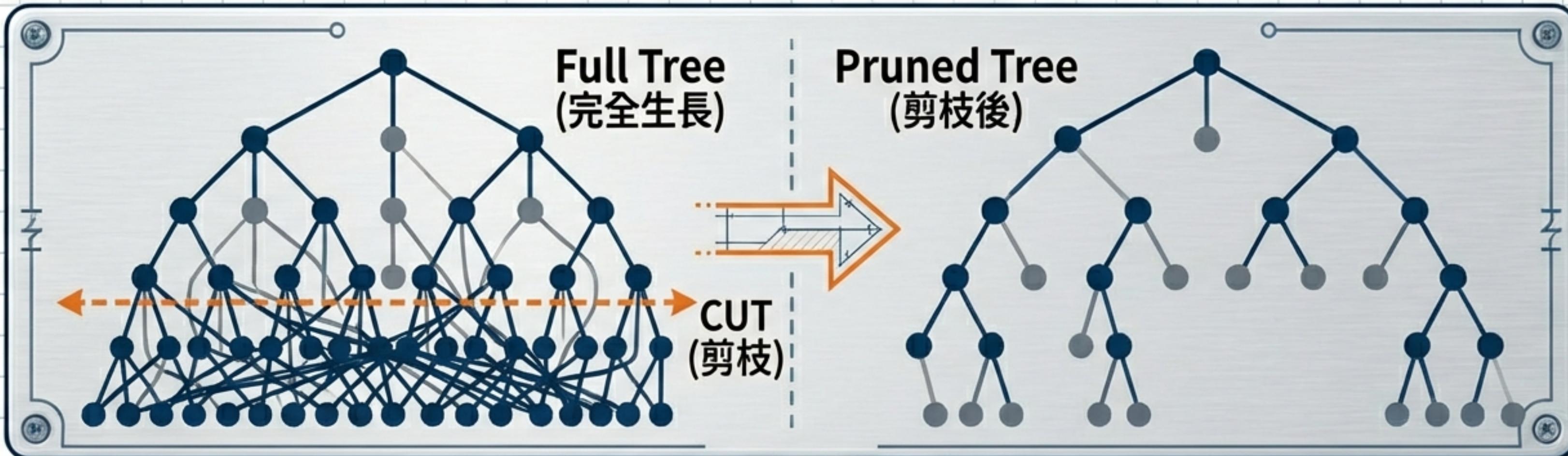


⚠ Noto Sans TC

建議：這是平衡 '偏差 (Bias)' 與 '變異 (Variance)' 的關鍵控制閥。

Noto Sans TC 剪枝策略 (Pruning Strategies)

去蕪存菁，提升模型泛化能力



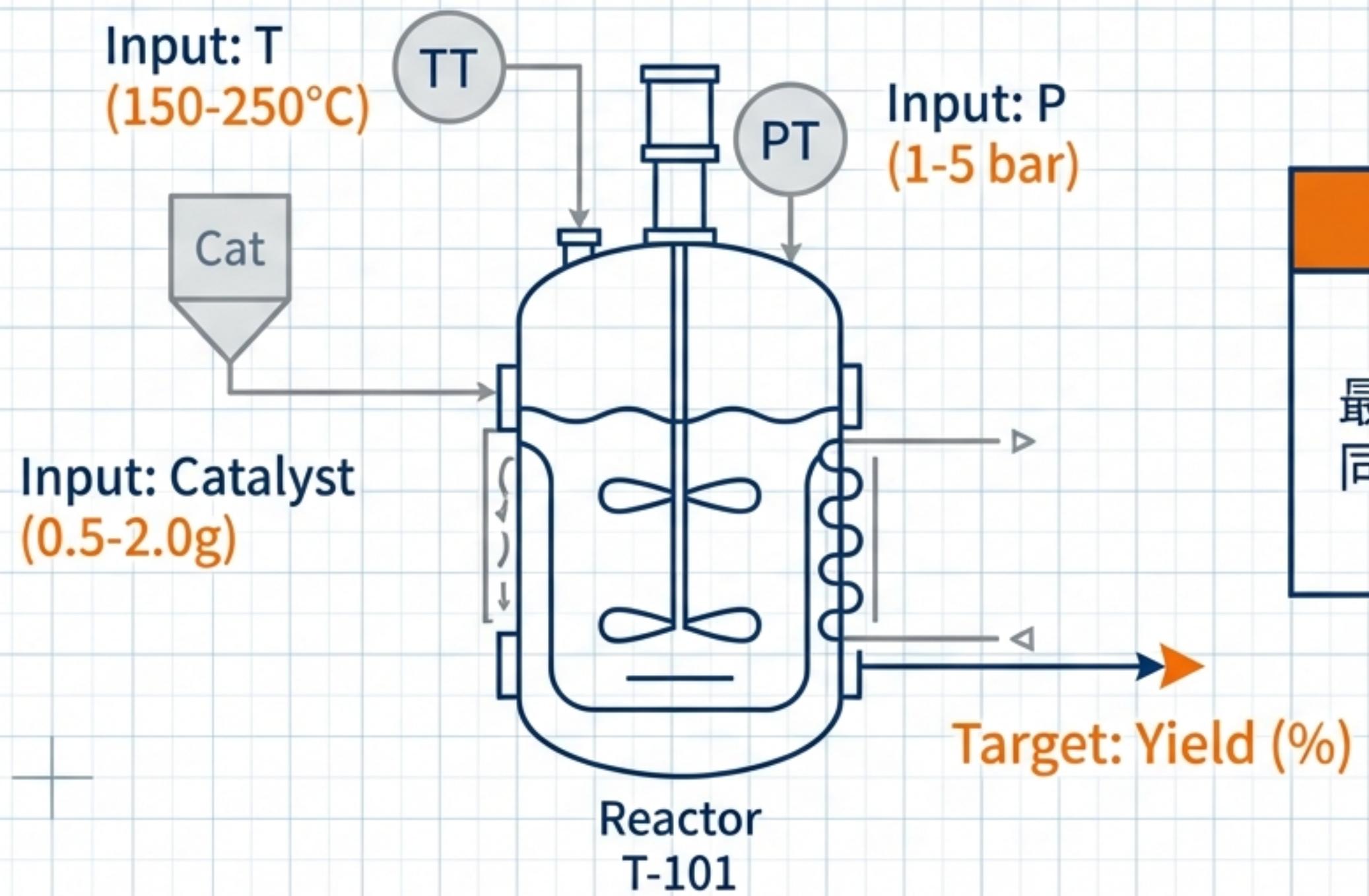
預剪枝 (Pre-Pruning)

- 在生長過程中提前停止
- 方法：限制 `max_depth`
- 優點：快 / 缺點：可能視短 (Short-sighted)

後剪枝 (Post-Pruning / CCP)

- 先讓樹完全生長，再根據 “成本複雜度” 切除無效子樹
- 優點：更穩健 / 缺點：計算量稍大

化工案例：反應器產率優化



No戰 Sans TC

最大化產率 (Maximize Yield)
同時最小化成本。

• 模型表現與診斷 (Model Performance & Diagnosis)

Model A (無限制樹 Unconstrained)

Train R²: **1.0000**

Test R²: **0.9013**



過擬合 (Overfit)

Model B (優化樹 Optimized - Depth 7)

Train R²: **0.9736**

Test R²: **0.8759**

MAE: **1.73%**



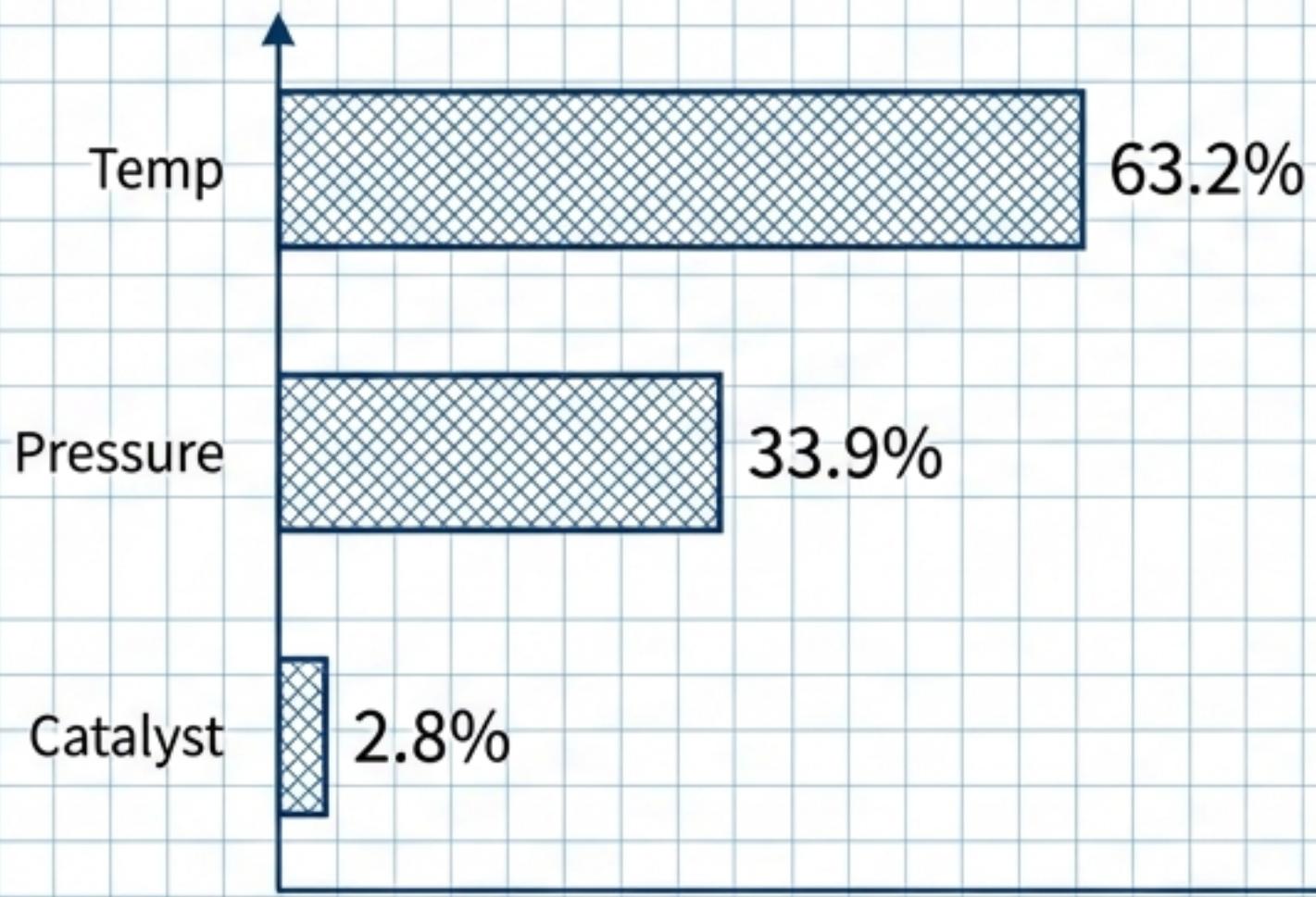
穩健 (Robust)

關鍵洞察 (Key Insight): 犥牲一點訓練準確度，換取對未知數據的可靠預測。

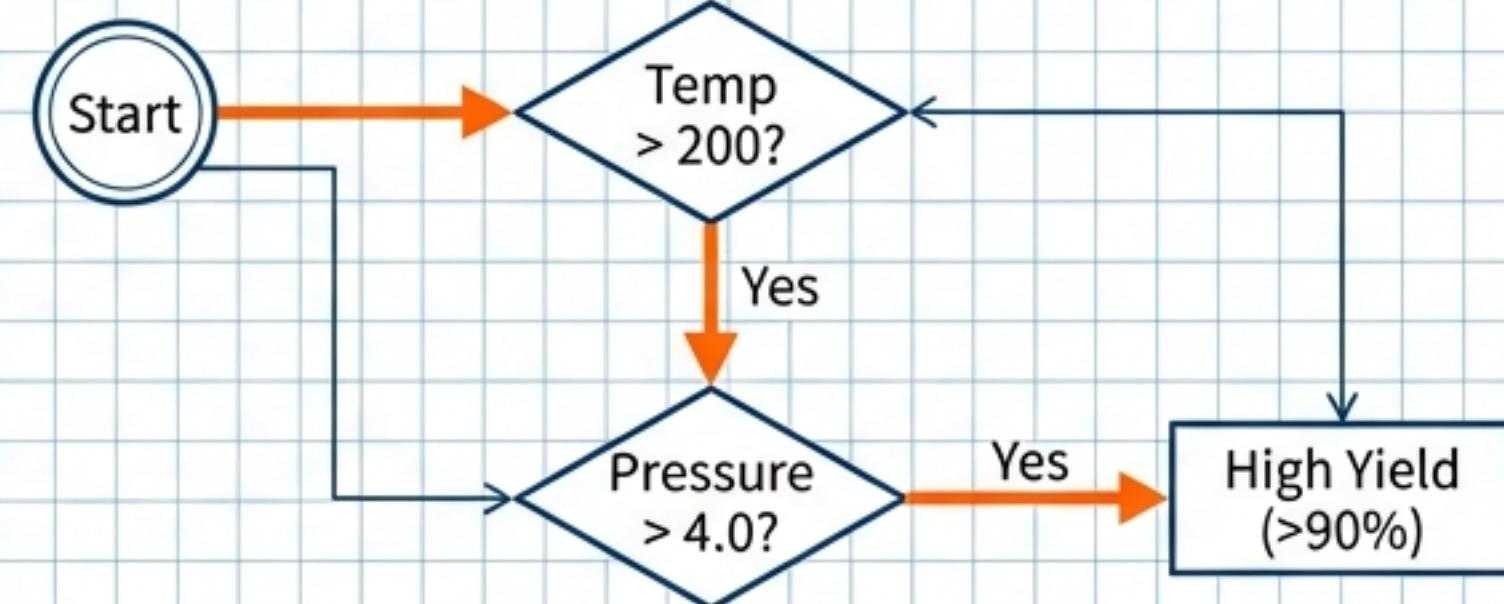
解讀黑箱：特徵重要性與規則 (Interpretation)



特徵重要性 (Feature Importance)



Golden Path (Winning Rule)

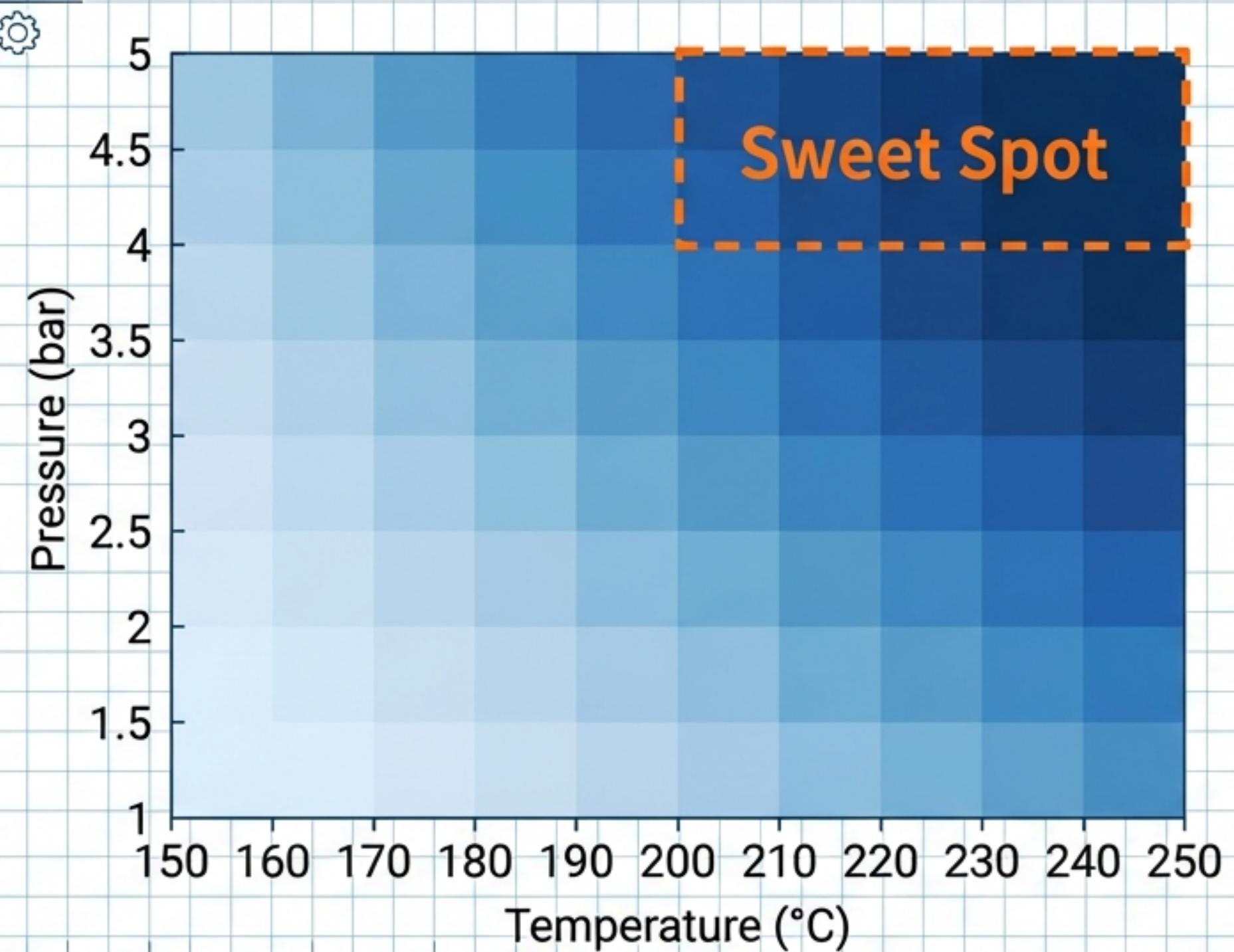


解讀工程規則 (Decoded Engineering Rules)

1. IF $\text{Temp} < 190^{\circ}\text{C}$ \rightarrow Low Yield (Arrhenius Limit)
2. IF $\text{Temp} > 200^{\circ}\text{C}$ AND $\text{Pressure} > 4.0 \text{ bar}$ \rightarrow High Yield ($>90\%$)

物理意義 (Physical Meaning) : 驗證了溫度的主導性與壓力的協同效應。

最佳操作條件搜尋 (Finding the Optimal Operating Window)



操作建議 (Actionable Advice):

- Temperature: **200 - 210°C**
- Pressure: **4.0 - 4.8 bar**
- Cost Saving: 催化劑可降至 0.78g 而不犧牲產率。

Warning: 避免極高溫 ($>235^{\circ}\text{C}$) 以防副反應。

模型選擇指南 (Model Selection Guide)

線性回歸 (Linear Regression)	決策樹 (Decision Tree)	隨機森林 (Random Forest)
<ul style="list-style-type: none">簡單、可外推 (Simple, Extrapolates)無法處理非線性 (Fails on Non-linear)	<ul style="list-style-type: none">自動處理非線性/交互作用 (Automatically handles non-linearity/interaction)可解釋性高 (Explainable)無法外推，階梯狀預測 (Cannot extrapolate, step-function predictions)	<ul style="list-style-type: none">準確度最高 (High Accuracy)魯棒性強 (Robust)黑箱 (Black Box)

建議：研發探索階段用 決策樹，追求極致準確度用 隨機森林。

結語：您是未來的定義者 (Conclusion)



物理意義 (Physics):
樹模型能捕捉數據中的物理非線性。



工程控制 (Control):
必須進行剪枝以確保安全與魯棒性。



價值 (Value):
真正的價值在於“解釋”(Feature Importance)。

下一步 (Next Step): 進階集成方法 — 隨機森林 (Random Forest)

AI 不會取代化工工程師，但懂得使用 AI 的化工工程師將取代不懂的人。