Gradient Boosting 模型參數解說 (Orange / Scikit-learn)

編撰：屏東大學 周國華老師 (與 Google Gemini 共筆) 2025/11/25

這張設定圖控制了機器學習中最強大的分類演算法之一：**梯度提升 (Gradient Boosting)**。

區塊 (Section)	參數名稱 (Parameter)	設定值 (Value)	技術意義 (Technical Significance)	教學重點 (Pedagogical Focus)
Basic Properties	**Number of trees** (樹的數量)	**100**	決定模型中**弱學習器** (即單一決策樹) 的總數量，也是模型迭代修正的次數。	這是模型複雜度的主要控制項。數量越多，模型越強大，但運算時間也越長。 (數量 100 是一個速度與效果的良好平衡點。)
	Learning rate (學習率)	**0.100**	控制每棵新樹在整體模型中的權重，即每次修正的幅度。	**「修正的謹慎度」**。低學習率 (如 0.1) 能使模型穩健收斂，不易被單一雜訊點帶偏，有助於提高泛化能力，但需要更多的樹 (即更高的「Number of trees」) 來彌補。
	Replicable training (可重複訓練)	**已勾選**	固定模型訓練時的隨機亂數種子 (Random Seed)。	**「結果的可靠性」**。對於需要驗證和審計的應用 (如逃漏稅分析)，此選項必須勾選，確保每次執行結果完全相同。
Growth Control	**Limit depth of individual trees** (限制單棵樹的深度)	**3**	控制每一棵單獨的決策樹最多能分岔的層數。	**「單一規則的複雜度」**。Gradient Boosting 偏好使用淺樹 (弱學習器)。深度 3 意味著單一判斷邏輯最多由三個特徵組合而成，這有助於避免過度擬合。

區塊 (Section)	參數名稱 (Parameter)	設定值 (Value)	技術意義 (Technical Significance)	教學重點 (Pedagogical Focus)
	Do not split subsets smaller than (不分割小於此數量的子集)	**2**	當一個節點只剩下 2 個樣本時，停止繼續分割。	**「防止過度擬合的機制」**。防止模型為了完美解釋極少數的邊緣樣本 (Outliers) 而建立不具泛化性的規則。
Subsampling	**Fraction of training instances** (訓練實例的比例)	**1.00**	每輪訓練中，用於建立新樹的樣本比例。	**「資料使用率」**。設定為 1.00 代表每次訓練都使用所有 500 筆資料。在樣本數適中時，使用全部資料是標準做法。

教學總結

這套配置 (100 棵樹、深度 3、學習率 0.1) 是一種經典的**「穩健集成學習」**策略。它確保了模型訓練過程的穩定性、結果的可驗證性，並在您的 500 筆資料上成功展示了強大的預測能力 (AUC 0.925)。