# 📘 第八章：分類（Classification）

## 📖 章節簡介與補充重點

本章介紹分類任務的基本概念與常用演算法，包括 KNN（K-Nearest Neighbors）、決策樹（Decision Tree）、隨機森林（Random Forest）等。分類是監督式學習中最常見的任務之一，其目標是從已知標記資料中學習分類邊界，進而對新資料進行預測。  
  
此外，本章也強調模型訓練與評估的技巧，例如混淆矩陣、精確率（Precision）、召回率（Recall）、F1 分數（F1 Score）等，並探討 Underfitting 與 Overfitting 的問題與解法。

## 🧠 名詞解釋（Term Explanation）

分類（Classification）  
中文：將輸入資料分為離散類別的一種監督式學習。  
英文：Supervised learning that categorizes input data into discrete classes.

KNN（K-Nearest Neighbors）  
中文：將新樣本歸類為 k 個鄰近樣本中最多的類別。  
英文：Classifies a sample based on the majority class among its k nearest neighbors.

決策樹（Decision Tree）  
中文：透過資料特徵不斷分裂節點形成樹狀結構來進行分類。  
英文：A tree-like model that splits data based on features to classify outcomes.

隨機森林（Random Forest）  
中文：集成多棵決策樹進行預測以提升準確率與穩定性。  
英文：An ensemble of decision trees that improves prediction accuracy and robustness.

混淆矩陣（Confusion Matrix）  
中文：列出預測與實際類別的交叉分布，用以評估分類模型效能。  
英文：A matrix showing the relationship between actual and predicted labels.

精確率（Precision）  
中文：真正例在所有被預測為正例中的比例。  
英文：The ratio of true positives to all predicted positives.

召回率（Recall）  
中文：真正例在所有實際正例中的比例。  
英文：The ratio of true positives to all actual positives.

F1 分數  
中文：Precision 與 Recall 的調和平均數，用來綜合評估模型效能。  
英文：The harmonic mean of precision and recall for balanced evaluation.

Overfitting  
中文：模型過度擬合訓練資料，對測試資料表現差。  
英文：Model fits training data too closely and performs poorly on unseen data.

Underfitting  
中文：模型無法擬合訓練資料，準確率低。  
英文：Model is too simple to capture the underlying patterns of the data.

## 📚 名詞比較（Term Comparison）

📌 KNN vs. 決策樹  
中文：KNN 是基於鄰近樣本，無訓練過程，預測時間長；決策樹有訓練過程，預測快但容易過擬合。  
英文：KNN is instance-based, slow at prediction; Decision Tree builds a model, fast prediction but prone to overfitting.  
  
📌 Precision vs. Recall  
中文：Precision 著重於預測正確率；Recall 著重於找出所有正例。若應用在醫療領域，通常更重視 Recall。  
英文：Precision focuses on prediction accuracy; Recall emphasizes capturing all positives. In medical applications, Recall is often prioritized.  
  
📌 Overfitting vs. Underfitting  
中文：Overfitting 表示模型太複雜，Underfitting 表示模型太簡單。需透過調參與交叉驗證避免。  
英文：Overfitting implies overly complex model; Underfitting means too simple. Use tuning and cross-validation to prevent both.

## 📝 模擬出題與中英文詳解（包含解釋性問題）

Q1: Q1. 為什麼 KNN 預測時需要儲存整個訓練資料集？它有什麼優缺點？  
答：因為 KNN 是記憶式學習（lazy learner），沒有明確的訓練過程，預測時需計算所有資料距離，優點是簡單但效率低。  
KNN is a memory-based method (lazy learner) that stores all training data and computes distances at prediction. It's simple but slow.

Q2: Q2. 為何隨機森林能有效解決決策樹的過擬合問題？其做法為何？  
答：隨機森林透過多棵不同資料與特徵的決策樹進行集成，並使用多數投票決策，增加模型穩定性與泛化能力。  
Random Forest reduces overfitting by combining multiple decision trees with random sampling and majority voting.

Q3: Q3. Precision 與 Recall 何時可能產生衝突？如何透過 F1 分數解決？  
答：當預測正例數過多時，Recall 高但 Precision 低；反之亦然。F1 分數平衡兩者權重，適合整體評估。  
When predicting many positives, Recall may be high but Precision low, and vice versa. F1 Score balances both for fair evaluation.

Q4: Q4. 如何透過混淆矩陣評估分類模型的效能？請解釋 TP, FP, FN, TN 各代表什麼意思。  
答：混淆矩陣列出四種預測結果：TP 真正例、FP 假正例、FN 假負例、TN 真負例，可進一步計算 Precision 與 Recall。  
Confusion matrix displays TP (true positive), FP (false positive), FN (false negative), and TN (true negative), enabling metric computation.