113 學年度 元智大學 工業工程與管理學系 畢業專題

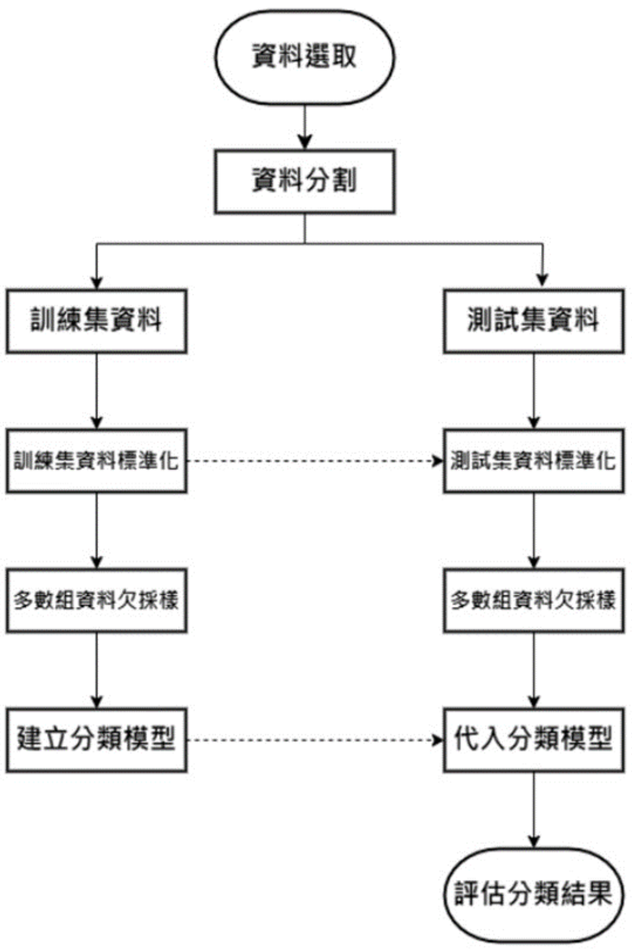
欠採樣於不平衡分類分析

指導教授:林真如　　學生:阮偉榤、李秉晁、陳奕瑾

研究動機與目的

在現代機器學習與生活當中，分類問題是常見的應用之一。例如，在寄送簡訊給消費者時，消費者是否回覆？一筆信用卡刷卡交易是否為盜刷？這些問題在大部分情況下都是陰性只有極少數是陽性，而這種不均衡現象就稱為「不平衡資料」。本研究將探討並比較多種處理不平衡資料的技術，並著重於欠採樣技術在不同應用場景中的效果與挑戰。我們將使用 Python 進行數據實驗與模型測試，從而驗證各種方法在不平衡資料下的分類性能，並為不平衡資料處理提供具體的實際建議。

研究方法

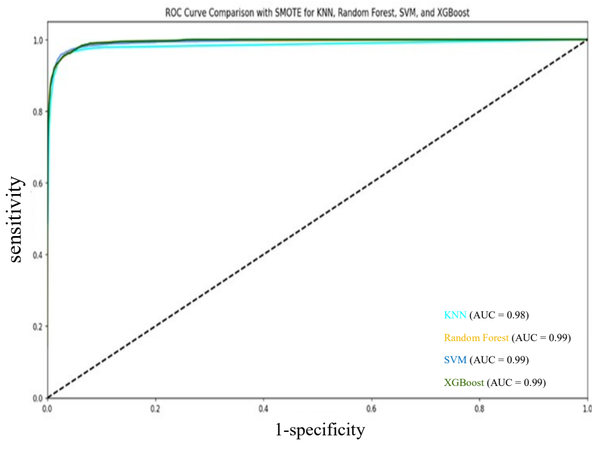


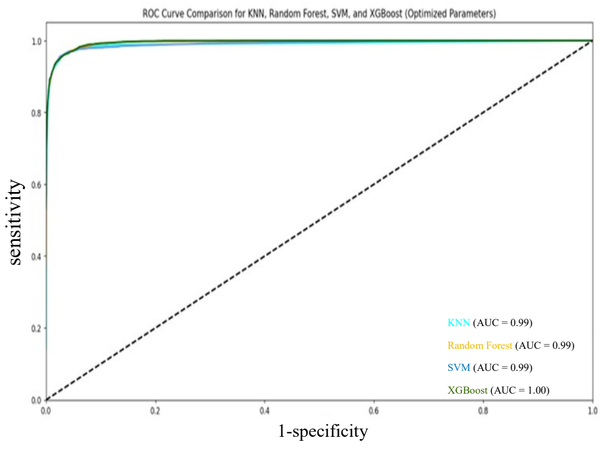
本研究使用的資料集包含七大類乾豆，且各類別樣本數量差異顯著，其中最大類別與最小類別的數量相差甚遠，這一特徵為模型訓練帶來挑戰。本研究分為三個階段進行：(1) 我們將最大類別的樣本數量下調至與最少類別相同進行單一欠採樣、(2)針對樣本數量最多的類別依照對樣本數量最少的類別進行削減、(3)所有類別同時進行欠採樣。透過這樣的比較，我們希望能更全面地理解不同欠採樣方案對模型預測績效及資料分佈結構的影響，並為實務中方法的選擇提供參考依據。

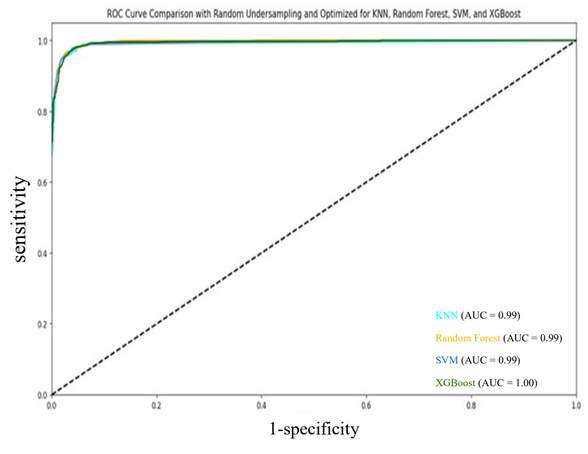
資料來源是使用 UCI Machine Learning Repository 中的 Dry Bean 數據集

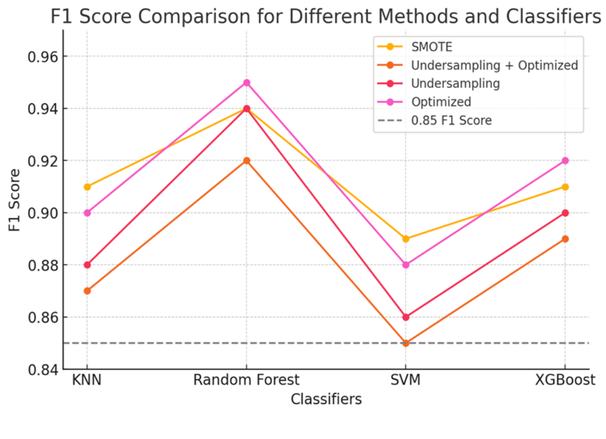
研究結果

本研究使用 ROC 曲線與 F1_Score 來評估分類模型表現，橫軸代表假陽性率，也可以理解為 1-specificity，而縱軸則是召回率(recall)或靈敏度(sensitivity)，表示模型正確預測正類別的比例而 AUC 是 ROC 曲線下的面積並利用 SMOTE、最佳化參數、最優化+隨機欠採樣分別對 KNN、Random Forest、SVM、XGBoost 進行實驗。



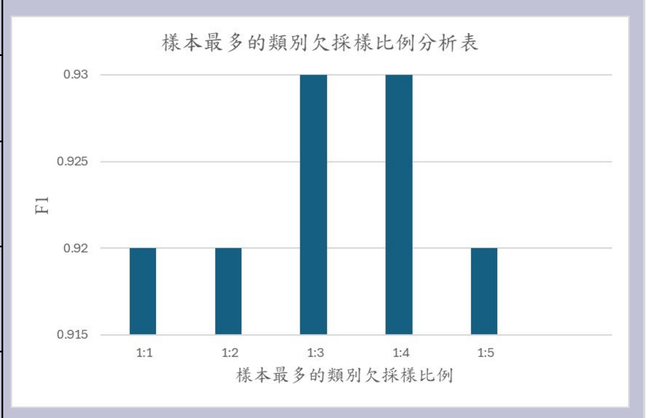




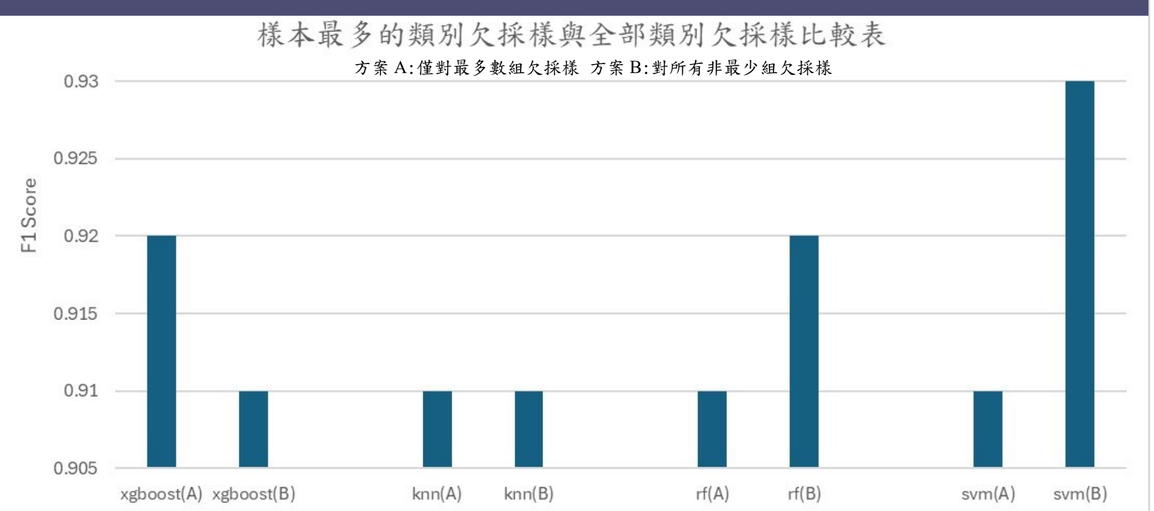


樣本最大類別欠採樣比例影響

組別	各組別比例數量				
樣本比例	1:1	1:2	1:3	1:4	1:5
最少樣本數量組別	417	417	417	417	417
最多樣本數量組別	417	834	1251	1668	2085
F1_Score	0.92	0.92	0.93	0.93	0.92



最大類別欠採樣與全域性欠採樣差異



結論

- 使用完全均衡的欠採樣策略能顯著提升多數分類模型的性能，尤其對於 Random Forest 和 SVM 效果顯著。
- 當數據不平衡較為嚴重時，建議選擇對不平衡度敏感的模型如 SVM，並結合均衡的欠採樣策略以達到最佳分類效果。
- 適度的欠採樣比例（1:3 至 1:4）可讓 XGBoost 達到最佳 F1_Score，保持多數類別的特徵並避免分類偏誤。