利用質譜儀定序偵測細菌抗藥性

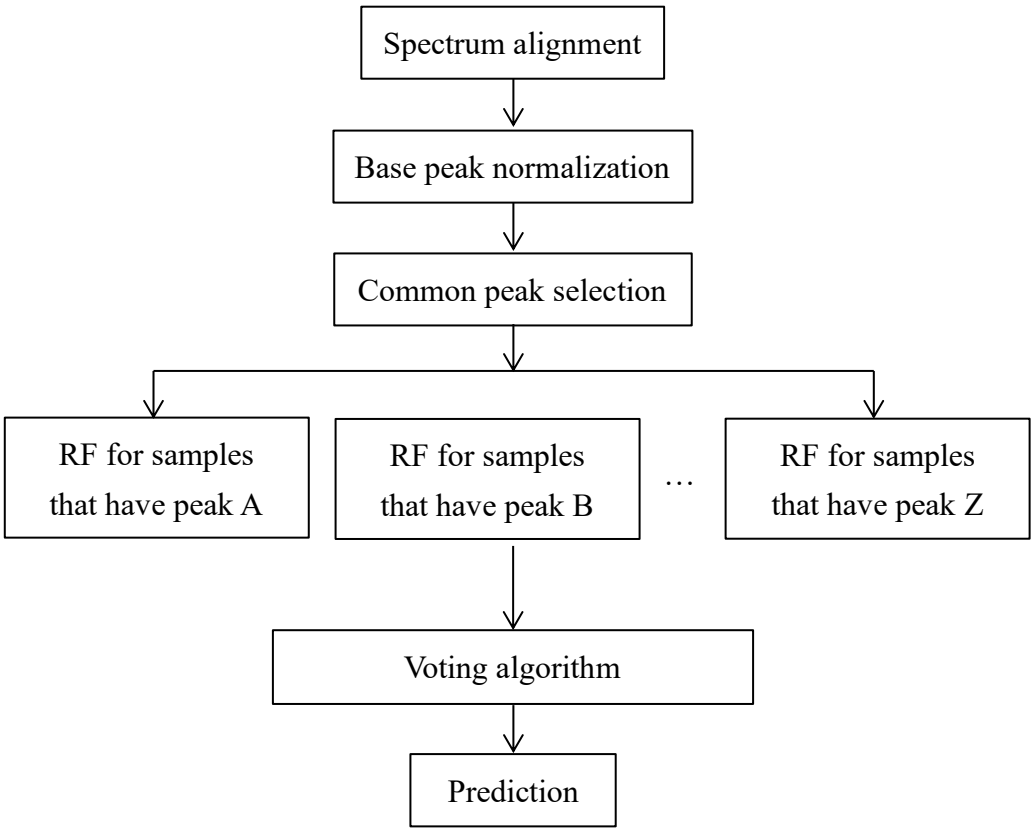
指導教授：莊坤達
專題成員：詹子毅
開發工具：Python, Sklearn
測試環境：macOS 10.14.6, Google Colab

一、簡介：

在住院照護中，院內感染是住院病患最常面對的風險之一。當病患不幸感染，反覆使用抗生素時常導致特定的病菌株對藥物產生抗藥性，進而惡化病情。由此可見，開發準確且快速的檢測病菌抗藥性的方法便顯得格外重要。目前醫院大多使用血液培養的方式檢測細菌對藥物的抗藥性，但是此方法需要 1 至 3 天的檢驗時間，於是臨床上開始尋找其他使用資料科學與數據分析檢測病菌抗藥性的替代方案。

此專題嘗試以集成學習為基礎，將細菌經過質譜儀分析所得的分子質量組成（可看做單一菌株的指紋）當作訓練資料，在經過頻譜校正等前處理後，透過集合多個對局部資料集具有高預測力的 Random Forest 模型，達到提昇 baseline 模型預測準確度的目標。

以下為模型架構圖：



二、測試結果：

	Decision Tree	Random Forest	AdaBoost	My Voting RFs
parameters	depth = 10	trees = 100	trees = 1000	
accuracy	0.639	0.677	0.700	**0.781**
R precision	0.838	0.728	0.700	**0.881**
R recall	0.366	0.591	0.724	0.660
R f1 score	0.510	0.652	0.712	0.754
S precision	0.582	0.641	0.700	0.720
S recall	0.926	0.769	0.674	0.908
S f1 score	0.715	0.699	0.686	0.803

（不具有抗藥性：S、具有抗藥性：R）

- 此資料集由 2477 筆資料組成，取出 20%作為測試資料，總計 496 筆。
- Voting RFs algorithm 總共訓練出 100 個 random forest，每個 random forest 包含 1000 棵樹，深度不限。
- 最後準確度由 baselines 的～69%提升至 78.1%，而 R precision 則提升至 88.1%，有 66%的抗藥性病株被模型偵測出來。
- 雖然模型有 S-biased 的傾向（如果找不到足以判定病株具有抗藥性的證據就傾向全部猜不抗藥），但是其實臨床上很難做到判斷 S 跟 R 都具有高準確度。例如新冠肺炎的快篩劑只能準確地篩選出陽性者，對於判別陰性的準確度並不高，所以能準確地幫助醫師排除確定具有抗藥性的病株才是在臨床上較有價值的成果。