# 紅白酒品質多分類問題

利用組合模型

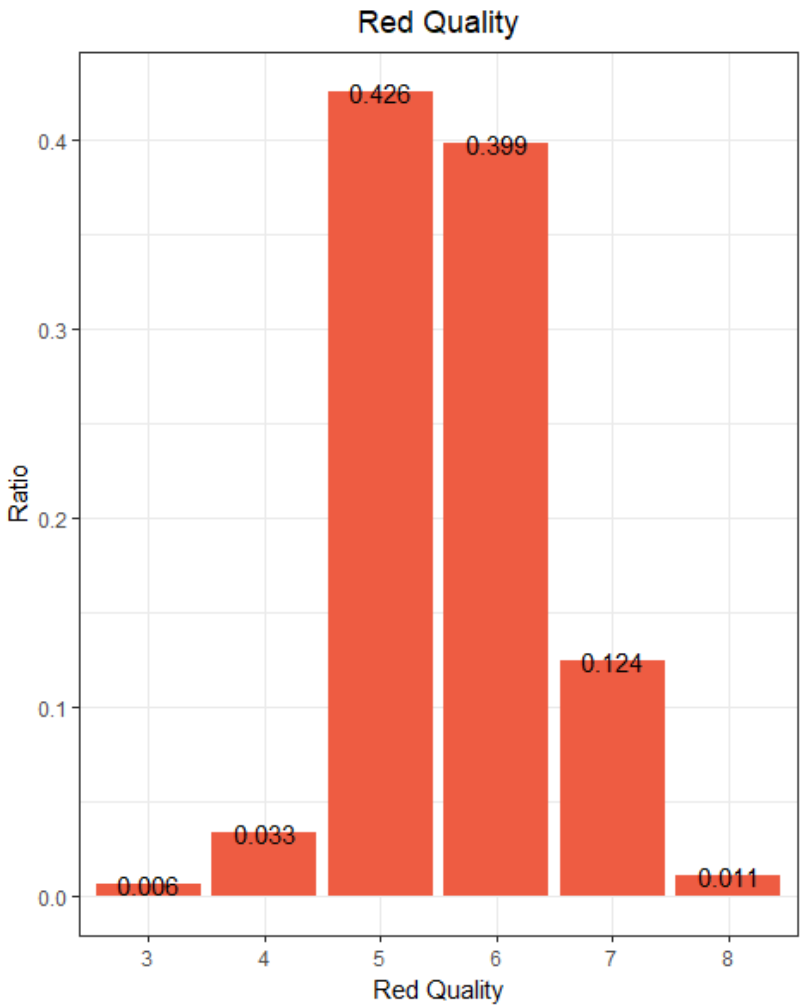
資管五 郭泰竹 102306037

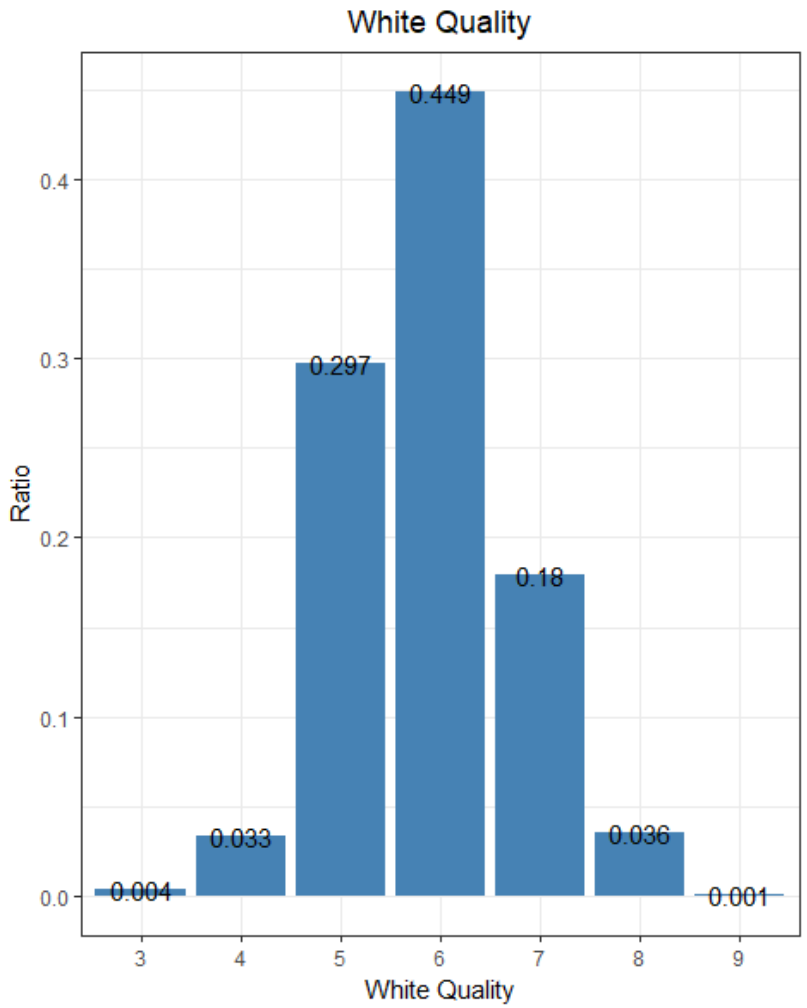
# Work Flow

Data Processing : 探索性分析 ➔ 資料合併 ➔ 資料切割

Data Modeling : 資料建模➔結果檢視➔組合模型➔成果解釋

# 探索性分析 - 紅白酒品質分布 ➔ 了解品質分布狀況



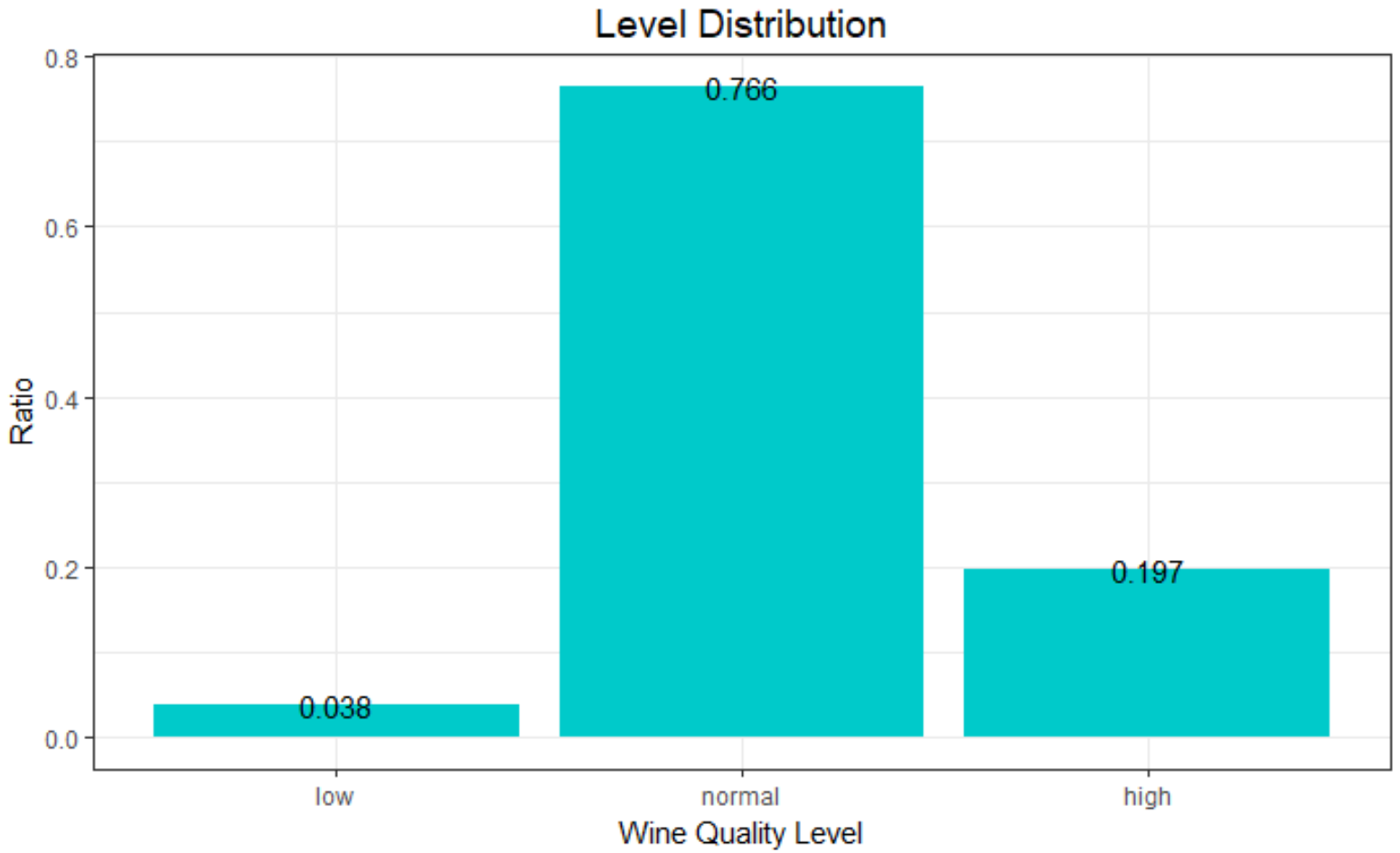

**說明**

紅酒分布以第五級、第六級最多

白酒分布以第六級為最多

➔等級分布上可能有些微差異

# 合併紅白酒資料 - 比例分析




**說明**

(1) 紅白酒資料合併後，比例上有明顯差異，比例約為1:20:5，在配模及切割訓練資料時須注意。

# 資料分割 - Train and Test set split

**Training Data**

6000筆
low:normal:high ➔228:4596:1176

**Testing Data**

497筆
low:normal:high ➔ 18:378:101

5-folds CV

**Training CV**

4800筆

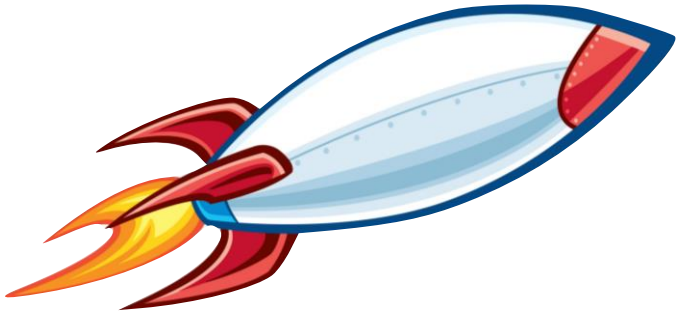
**Testing CV**

1200筆

**說明**

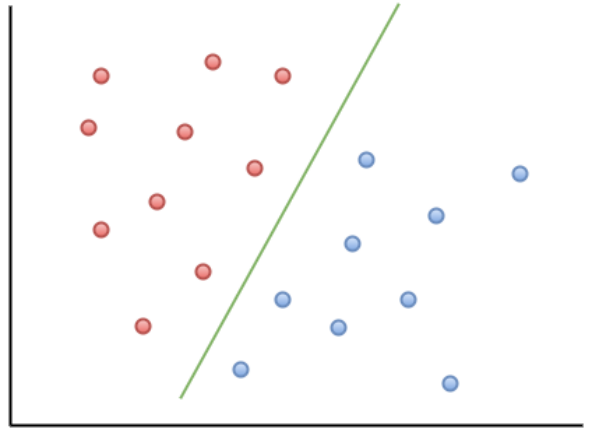
(1) 共有6497筆資料，取6000筆為training data，497筆為testing data。
(2) 分割時按照品質的比例切割，使train 和 test當中的比例相同。
(3) 在Training data中再切成Training CV 以及 Testing CV ，做 5-Fold CV，包含品質的比例也相同。

# 資料建模 - 使用模型



XGBoost

Random Forest



SVM

# 獨立資料建模 - Random Forest (original) 準確率 85.2%

```
$acc
[1] "0.852±0"

$recall
[1] "0.569±0.03"

$precision
[1] "0.805±0.03"
```

測試資料預測結果

| answer | pred_original low | normal | high |
|---|---|---|---|
| low | 3 | 15 | 0 |
| normal | 0 | 362 | 16 |
| high | 0 | 39 | 62 |

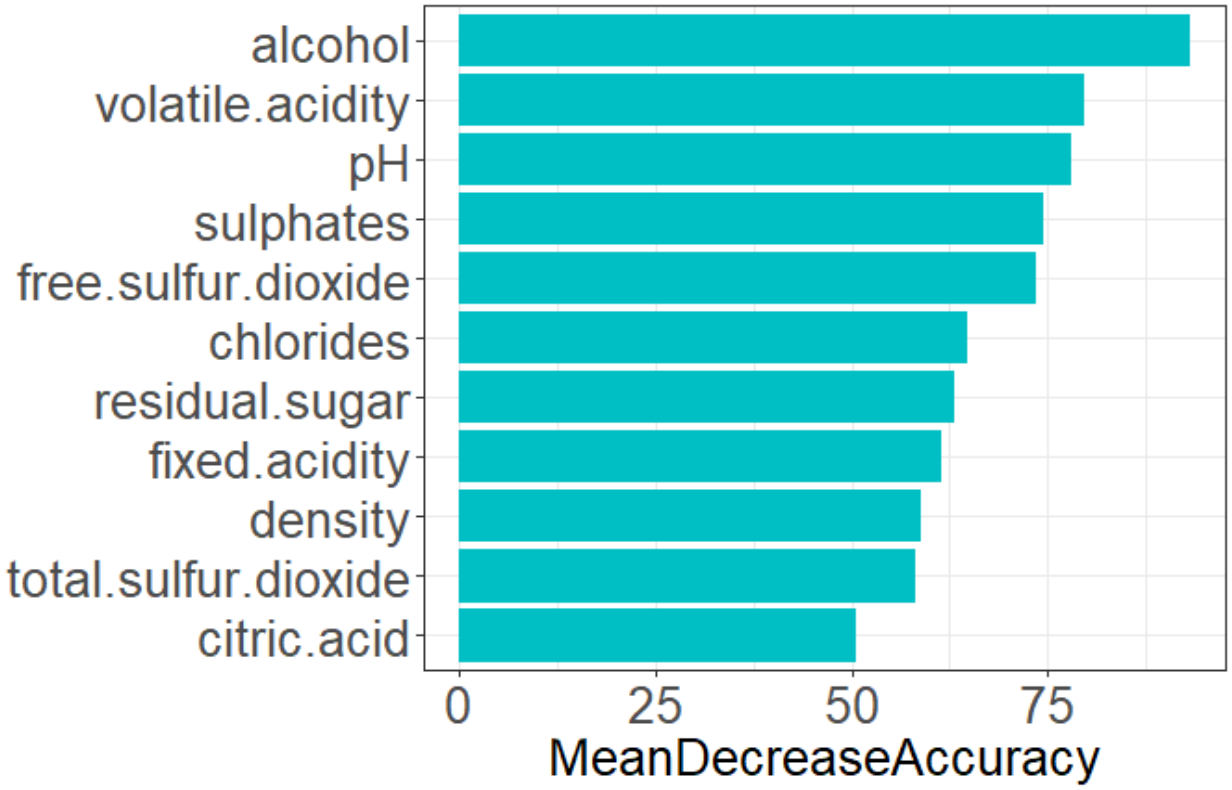
**說明**

(1) 訓練精準度(交互驗證)➔85.2%
(2) 訓練召回率(高品質)➔57.6%(所有高品質中，有幾個被猜中)
(3) 訓練精確率(高品質)➔80%(所有猜高品質的，猜中幾個)

**說明**

(1) 測試精準度➔85.9%
(2) 混淆矩陣 ➔low 召回率 = 16.7%
➔normal 召回率= 95.7%
➔high 召回率= 38.6%

# 變數解釋 - Importance plot




**說明**

**(1) Alcohol(酒精),volatile.acidity(揮發性酸度)對於品質分類模型有高度影響。**

**說明**

MDA值愈大，該特徵變數對於該模型的判別影響愈大

**(1) MeanDecreaseAccuracy ➔分別對每個特徵的值改變為隨機數，然後對所有樹計算單獨改變後的準確性平均差距，並利用標準差進行標準化。**

**簡單來說，MDA就是指 改變這個特徵的值 會造成多大影響。**

# 獨立資料建模 - SVM(original) 準確率 79.6%

```
$acc
[1] "0.796±0.01"

$recall
[1] "0.292±0.05"

$precision
[1] "0.679±0.03"
```

說明

(1) 訓練精準度(交互驗證)➔79.6%
(2) 訓練召回率(高品質)➔29.2%(所有高品質中，有幾個被猜中)
(3) 訓練精確率(高品質)➔67.9%(所有猜高品質的，猜中幾個)

測試資料預測結果

| answer | pred_svm low | normal | high |
|---|---|---|---|
| low | 0 | 18 | 0 |
| normal | 0 | 365 | 13 |
| high | 0 | 66 | 35 |

說明

(1) 測試精準度➔80.4%
(2) 混淆矩陣 ➔low 召回率 = 0%
➔normal 召回率= 96.5%
➔high 召回率 = 34.6%

# 獨立資料建模 - XGBoost(original) 準確率 82.8%

```
$acc
[1] "0.828±0.01"

$recall
[1] "0.509±0.02"

$precision
[1] "0.728±0.03"
```

測試資料預測結果

| answer | pred_xgb low | normal | high |
|---|---|---|---|
| low | 0 | 18 | 0 |
| normal | 1 | 359 | 18 |
| high | 0 | 55 | 46 |

**說明**

(1) 訓練精準度(交互驗證)➔82.8%

(2) 訓練召回率(高品質)➔50.9%(所有高品質中，有幾個被猜中)

(3) 訓練精確率(高品質)➔72.8%(所有猜高品質的，猜中幾個)

**說明**

(1) 測試精準度➔81.4%

(2) 混淆矩陣 ➔low 召回率 = 0%

➔normal = 94.9%

➔high = 45.5%

# 資料建模 - Ensemble Model(Stacking)

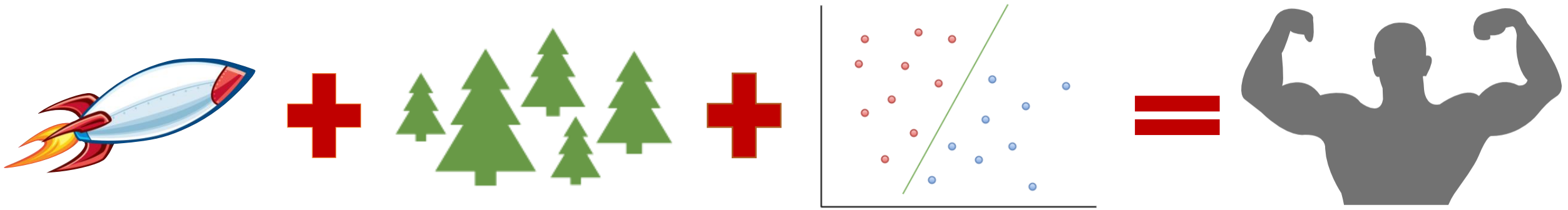
## 原理

(1) 利用集成式的方式，混合不同模型的資訊，製作新的模型。

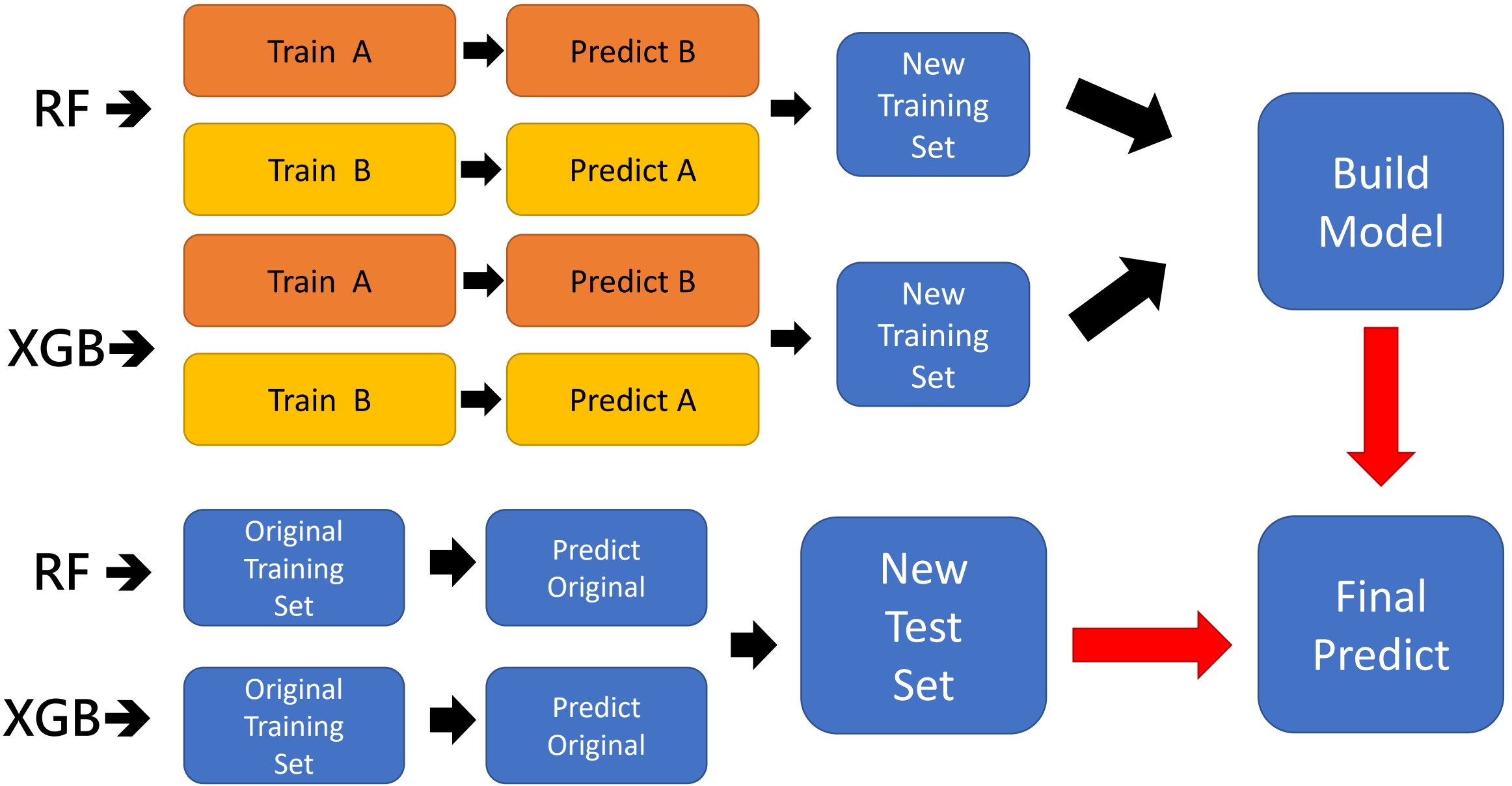
(2) 透過交互驗證預測的結果作為訓練的資料，去訓練新的模型，再利用模型預測的結果當作測試資料，做最後的預測。

## 優點

(1) 能夠彌補各模型在分類問題上的不足(線性、投票、增強式學習..等)，從更多角度看問題的感覺。



# 關鍵就是 ➔ 針對每一個模型做一次，並結合在一起產生新的



# 混合式資料建模 - Stacking(RF+XGBoost) 準確率 85.4%

```
$acc
[1] "0.854±0"

$recall
[1] "0.576±0.03"

$precision
[1] "0.811±0.03"
```

測試資料預測結果

| y_true | y_pred low | normal | high |
|---|---|---|---|
| low | 3 | 15 | 0 |
| normal | 0 | 364 | 14 |
| high | 0 | 41 | 60 |

**說明**

(1) 訓練精準度(交互驗證)➔85.4%

(2) 訓練召回率(高品質)➔57.6%(所有高品質中，有幾個被猜中)

(3) 訓練精確率(高品質)➔81.1%(所有猜高品質的，猜中幾個)

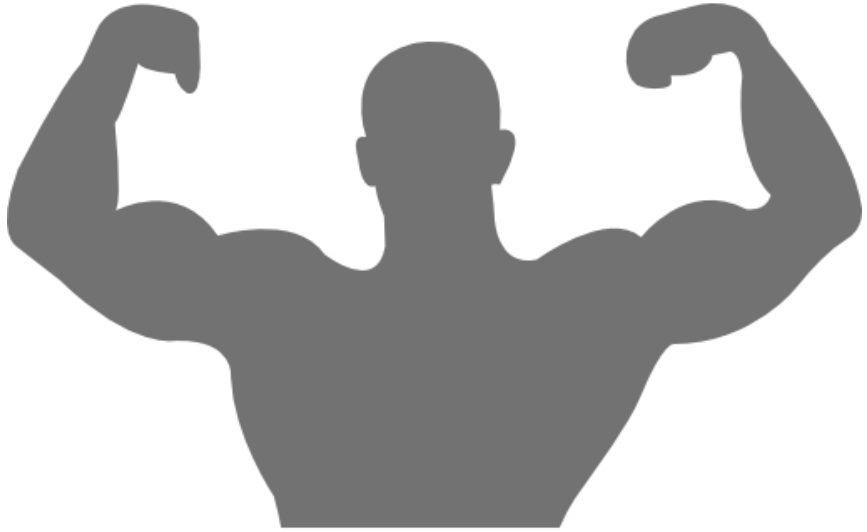
**說明**

(1) 測試精準度➔85.9%

(2) 混淆矩陣 ➔low 召回率 = 16.7%

➔normal 召回率= 96.2%

➔high 召回率= 59.4%



# Thanks for listening