dataset 搜集与整理 → Balance / InBalance

↓

排演算法建模 (model)

- KNN
- 貝氏演算法
- SVM
- deep learning

產生混淆矩陣

↳ 用測試資料建的

如果是用訓練資料，資料本身就"是"已知"的

算 accuracy 沒意義！

round ①, round ②

Training (75%) (70%)

隨机取

Testing (25%) (30%)

隨机取，做交叉驗證

為避免依賴特定資料產生的偏差

什麼是混淆矩陣？

↳ 用來判定一个模型表現的好、壞

Real \ Predict	Positive	Negative	
Positive	A TP	B FN	recall 召回率 $\frac{A}{A+B}$
Negative	C FP	D TN	specificity $\frac{D}{C+D}$
	Precision 精確率 $\frac{A}{A+C}$	negative predictive value $\frac{D}{B+D}$	Accuracy 正確率 $\frac{A+D}{A+B+C+D}$

多少正樣本被正確誤判

1. recall：真陽性比率
2. specificity：真陰性比率
3. 真陽性 Precision：$\frac{A}{A+C}$
4. 真陰性：$\frac{D}{B+D}$
5. error-rate (錯檢率)

多少負樣本被錯誤判別 $\Rightarrow \frac{B+C}{A+B+C+D}$

TP (True Positive)：正確預測正樣本

TN (True Negative)：正確預測負樣本

FP (False Positive)：錯誤預測正樣本

FN (False Negative)：錯誤預測負樣本

6. F1-measure (準確率&召回率的調和平均)

$= \frac{2A}{2A+b+c}$

Ex：判斷是否有 covid-19

TP：真的陽性且判斷陽性

TN：〃陰性 〃陰性

FP：真的陰性但判斷陽性

FN：〃陽性 〃陰性 * 嚴重

* overfitting：雖然能完美的符合測試"訓練資料"，

但調試得太過精確，在新的測試資料中含有更高的 error rate