三種訓練module

randomForest

1. 用*N*來表示訓練用例（樣本）的個數，*M*表示特徵數目。
2. 輸入特徵數目*m*，用於確定決策樹上一個節點的決策結果；其中*m*應遠小於*M*。
3. 從*N*個訓練用例（樣本）中以有放回抽樣的方式，取樣*N*次，形成一個訓練集（即bootstrap取樣），並用未抽到的用例（樣本）作預測，評估其誤差。
4. 對於每一個節點，隨機選擇*m*個特徵，決策樹上每個節點的決定都是基於這些特徵確定的。根據這m個特徵，計算其最佳的分裂方式。
5. 每棵樹都會完整成長而不會[剪枝](https://zh.wikipedia.org/w/index.php?title=%E5%89%AA%E6%9E%9D&action=edit&redlink=1)（Pruning，這有可能在建完一棵正常樹狀分類器後會被採用

KNN

support vector machine, SVM

**Pandas 提供的資料結構**

* **1.Series**：用來處理時間序列相關的資料(如感測器資料等)，主要為建立索引的一維陣列。
* **2.DataFrame**：用來處理結構化(Table like)的資料，有列索引與欄標籤的二維資料集，例如關聯式資料庫、CSV 等等。
* **3.Panel**：用來處理有資料及索引、列索引與欄標籤的三維資料集。

特徵選擇(Feature Slection)

前向選擇法(RFE)做特徵選擇，關於特徵選擇的方法我們有幾個選項，一是用單變數的Chi square、或是information gain，但RFE可以考慮到特徵之間的交互作用，缺點是需要較大的運算資源，這點由於我們的資料集比較少(titanic)而可以獲得緩解

參考

<https://www.itread01.com/content/1523270300.html>

最佳化理論???