資料探勘 Project2 報告

資訊三 E44051259 陳冠霖

前言：

本次的project非常有趣，我將要自己訂定一個規則，並觀察機器是否能透過dataset學習到我們所訂定的規則，本次project中，我共實驗了三種規則，並搭配不同的資料筆數，以及符合規則的數量佔全部數量的比數，等這兩個參數來調整不同的資料集，藉此觀察最終的結果，並透過畫圖的方式來確認機器是否真的有學到。(分別比較三種分類模型：決策樹,隨機森林,xgboost)

資料集介紹：

每一筆資料集都會有一個不同的真理，分別是線性（考試及格與否）的非線性（面積），以及一個我自訂的規則。

一：如果期中和期末考的平均超過60分，則label為1

二：feature1是長度，feature2是面積，如果面積為長度平方，則判斷為正 方形label為1

三：如果feature1是1且feature2是2則label就是1

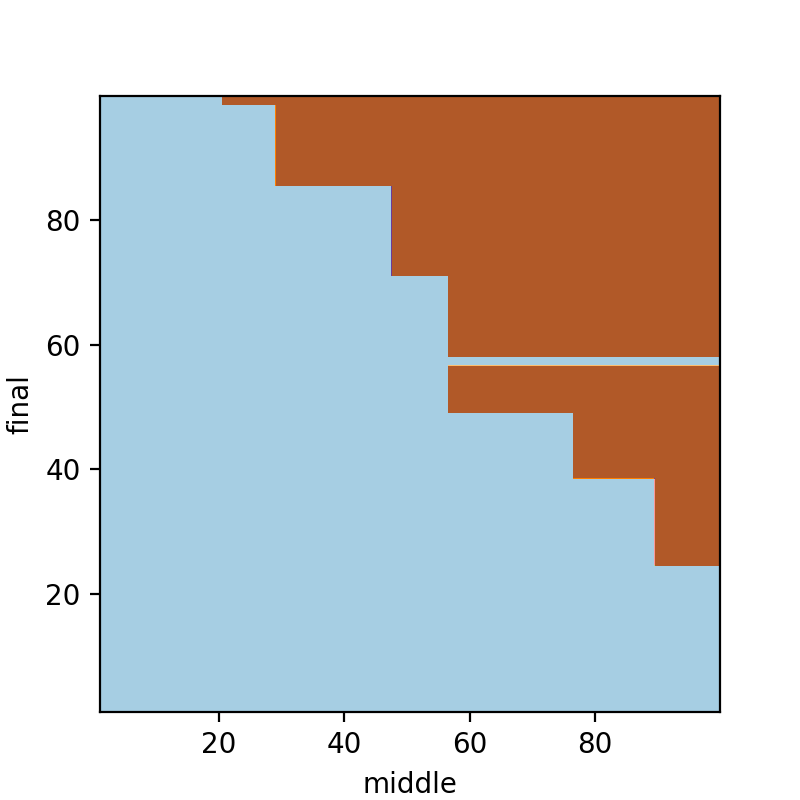
比較結果：

資料集一：

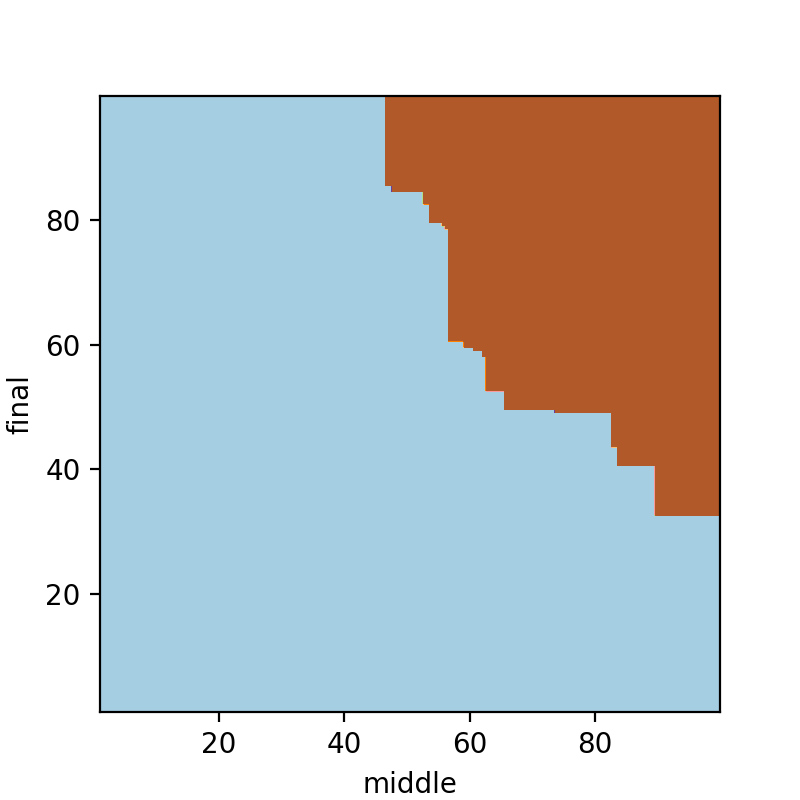
真理：如果期中和期末考的平均超過60分，則label為1

**資料筆數：500 符合真理的比數佔比：0.25**

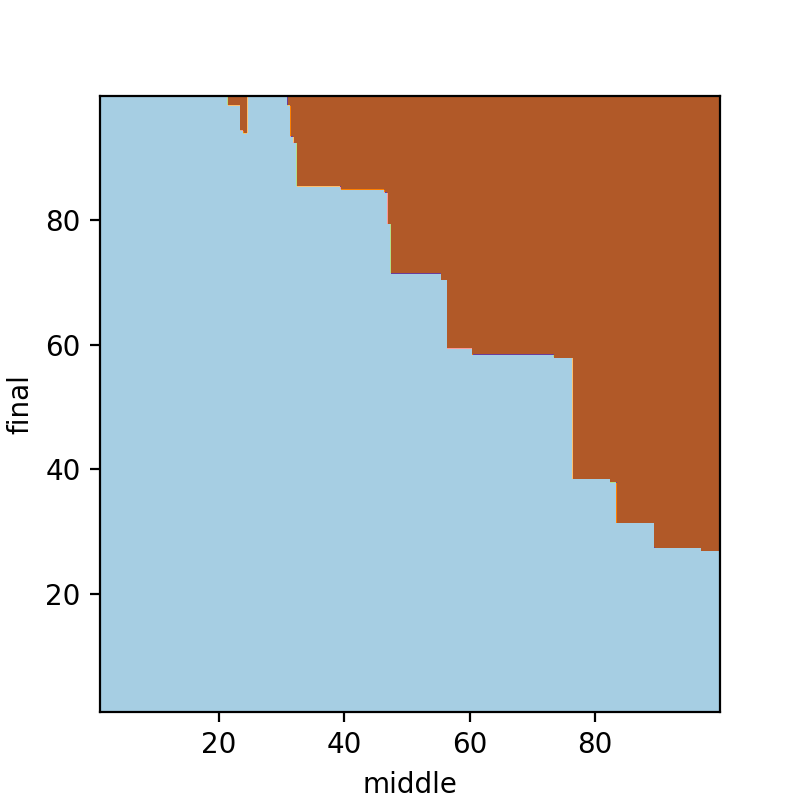
決策樹



隨機森林

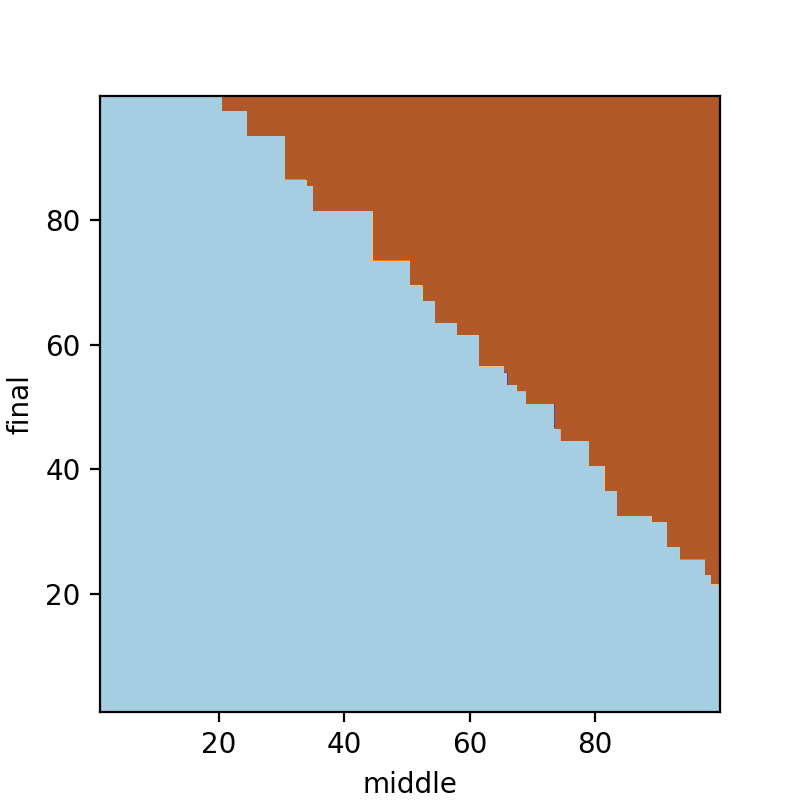


xgb

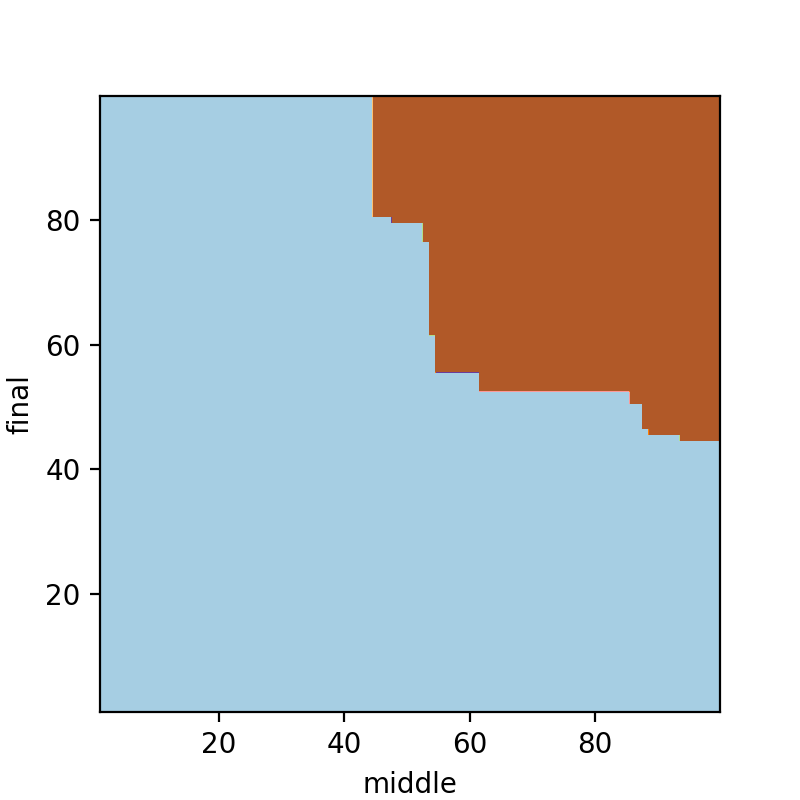


**資料筆數：5000 符合真理的比數佔比：0.25**

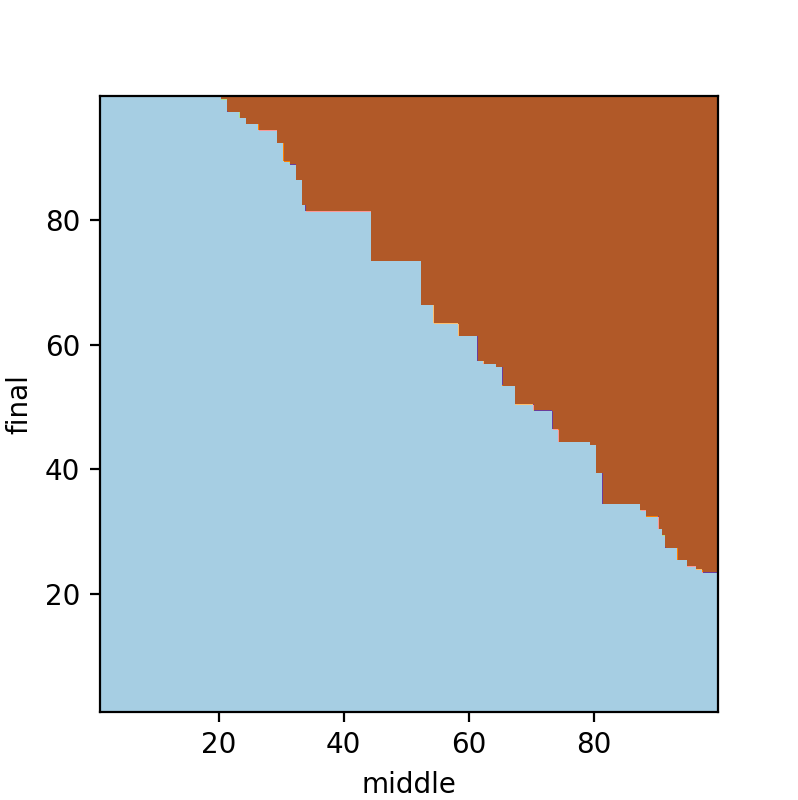
決策樹



隨機森林

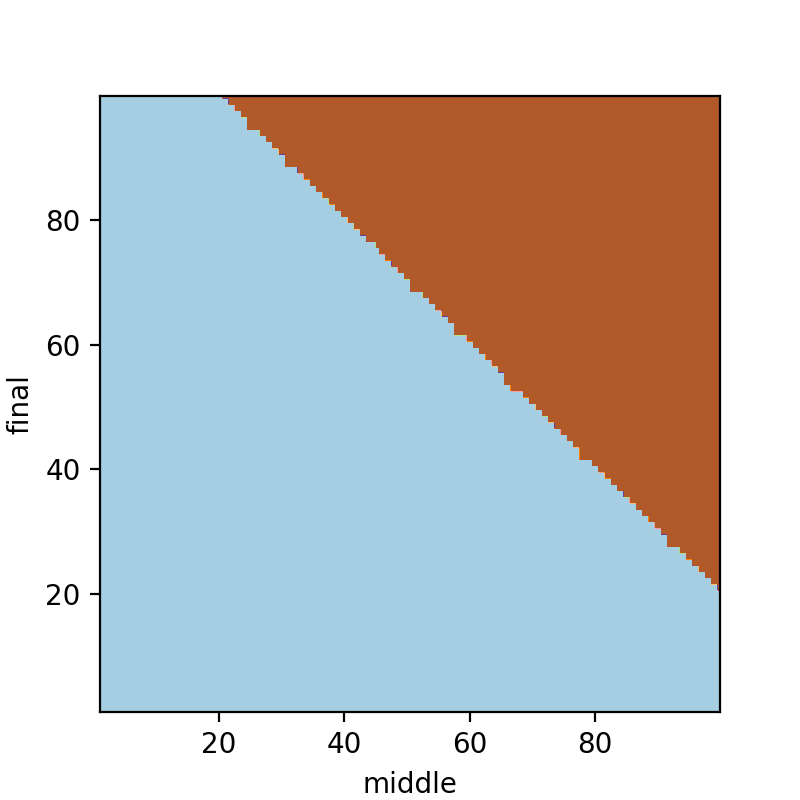


xgb

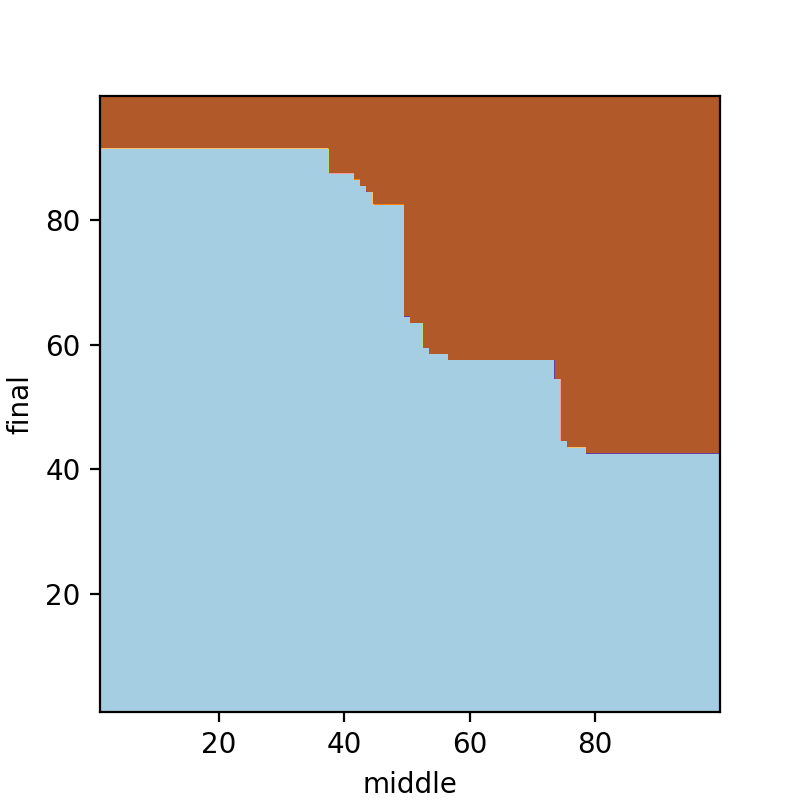


**資料筆數：50000 符合真理的比數佔比：0.5**

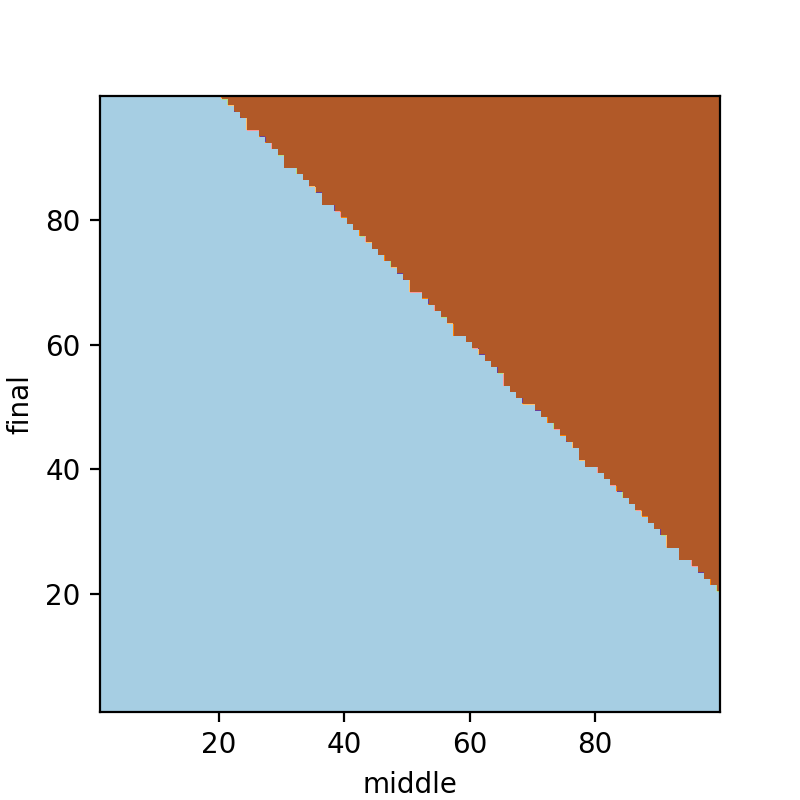
決策樹



隨機森林

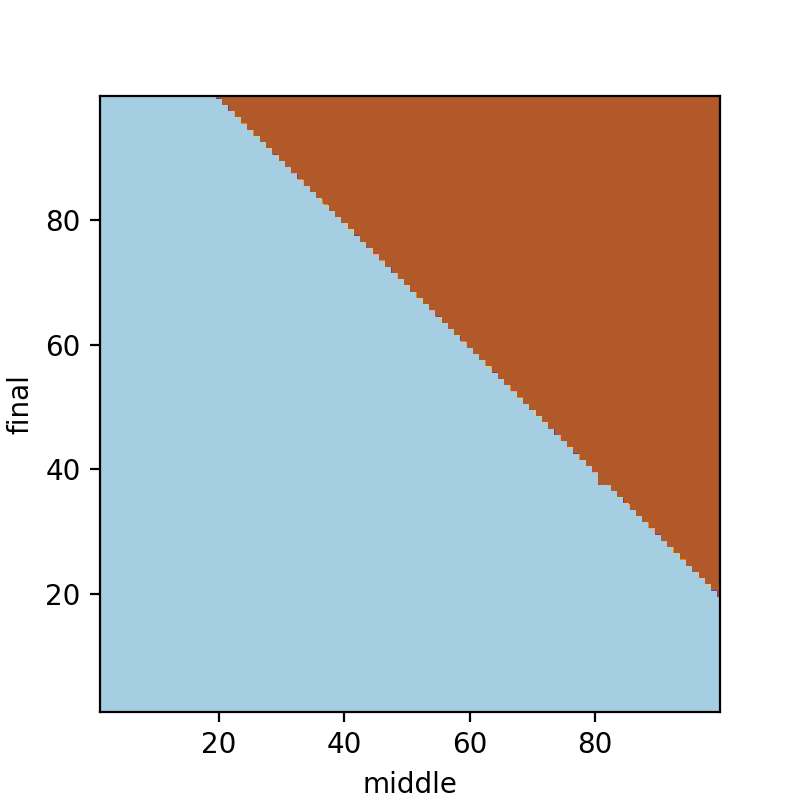


xgb

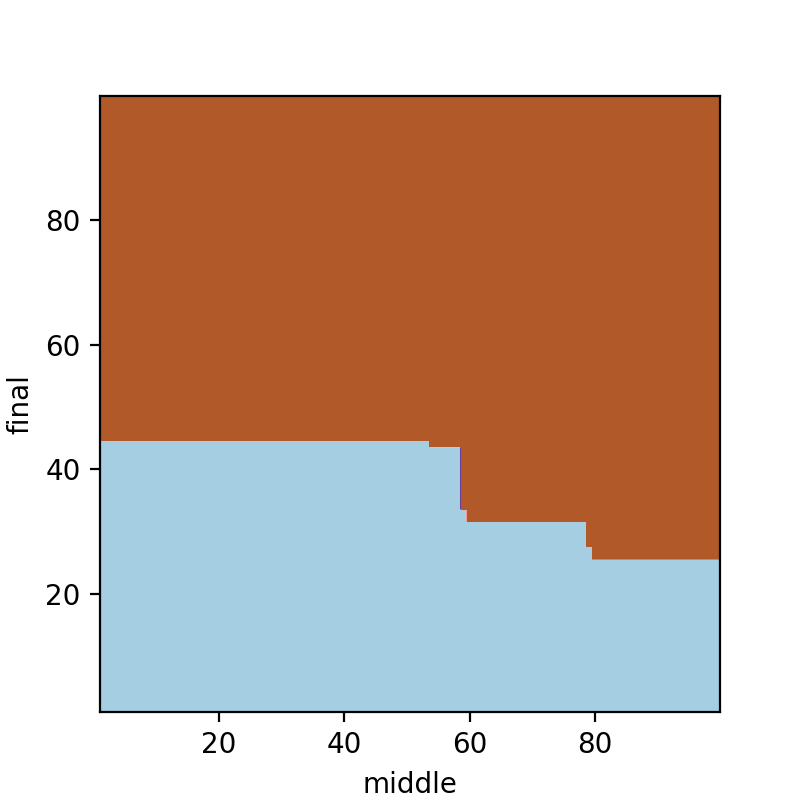


**資料筆數：50000 符合真理的比數佔比：3**

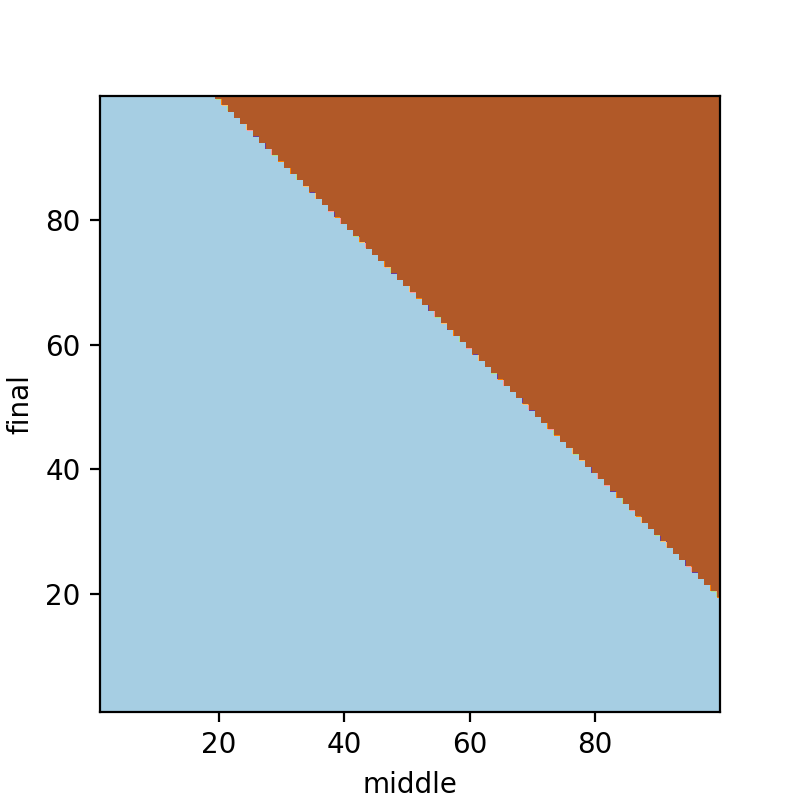
決策樹



隨機森林

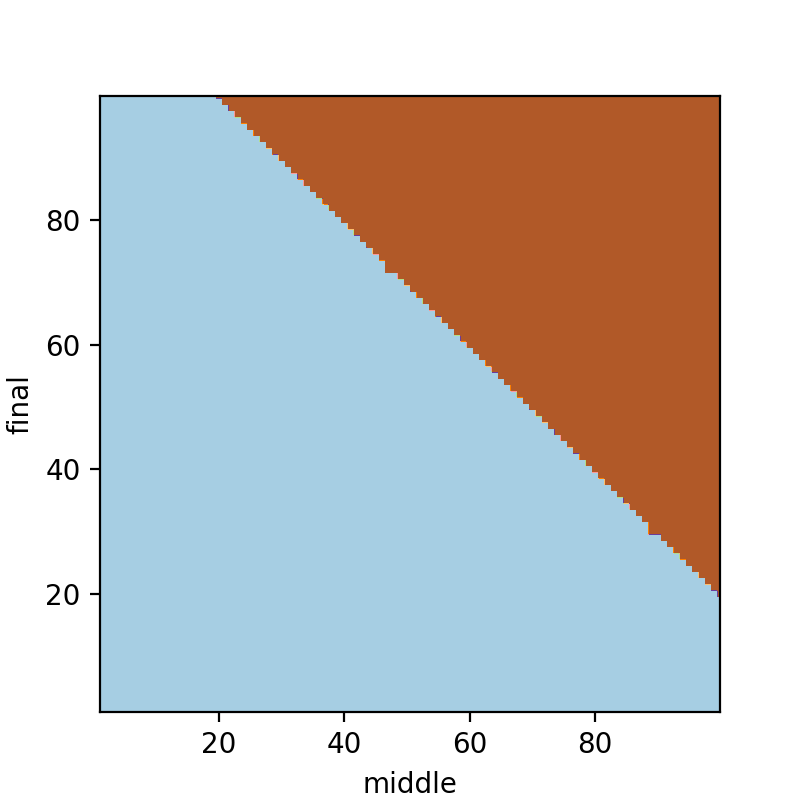


xgb

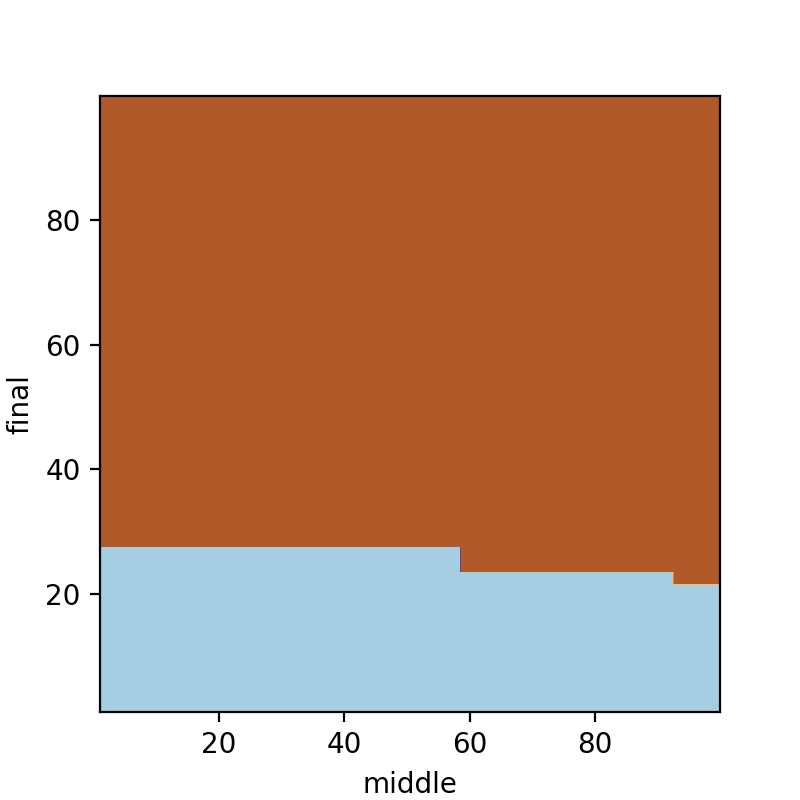


**資料筆數：50000 符合真理的比數佔比：9**

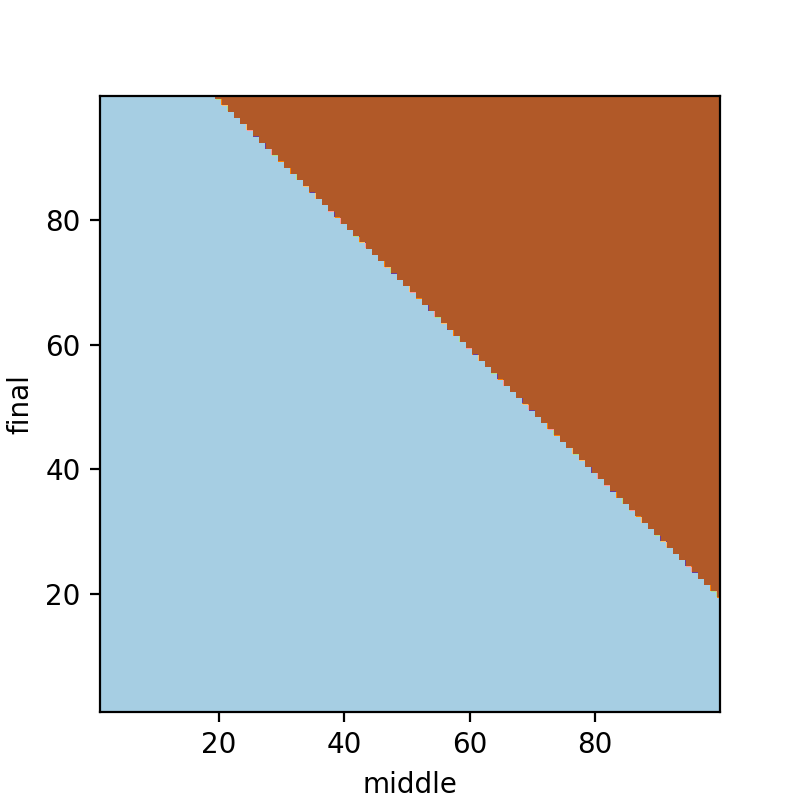
決策樹



隨機森林

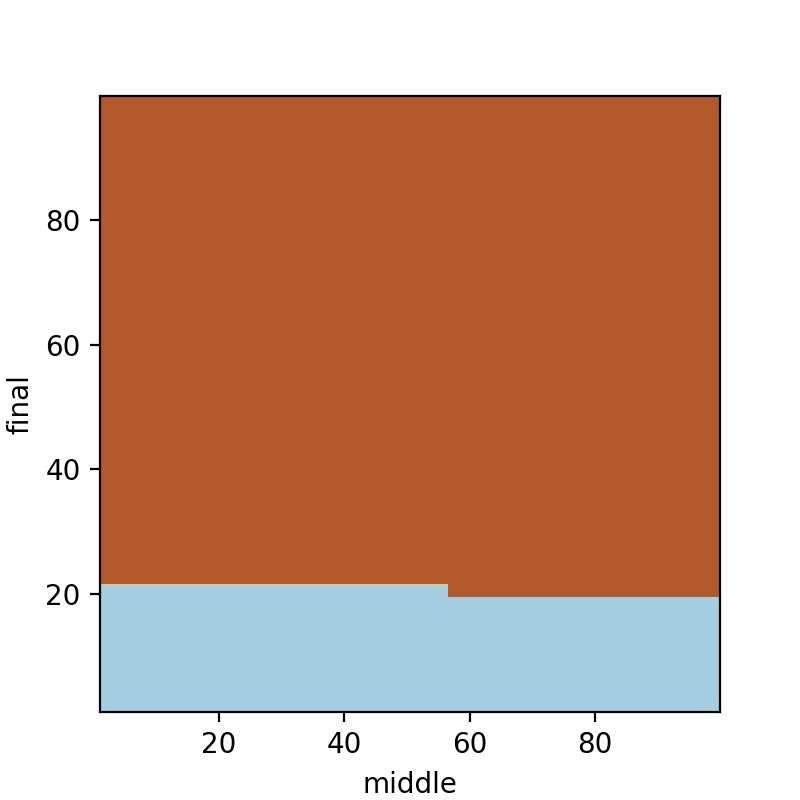


xgb



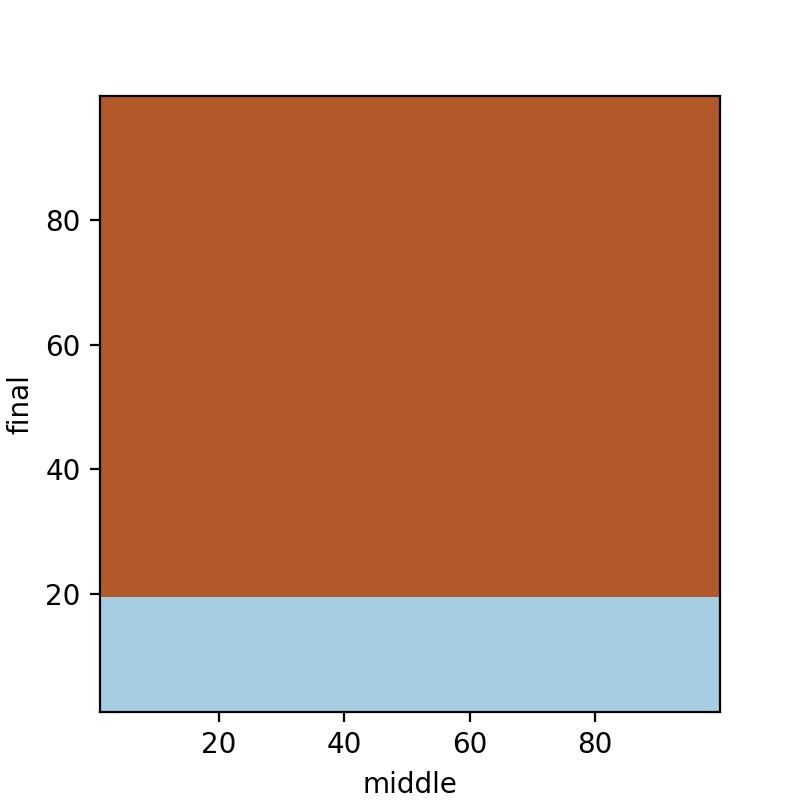
**資料筆數：50000 符合真理的比數佔比：30**

隨機森林



**資料筆數：50000 符合真理的比數佔比：90**

隨機森林

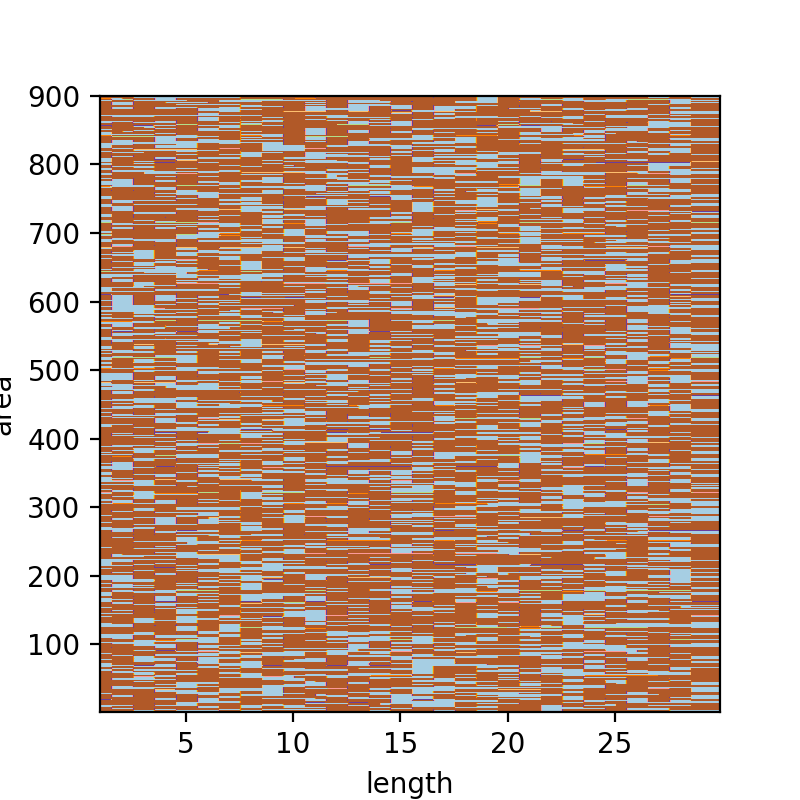


資料集二：

真理：feature1是長度，feature2是面積，如果面積為長度平方，則判斷為正 方形label為1

**資料筆數：50000 符合真理的比數佔比：3**

**決策樹**

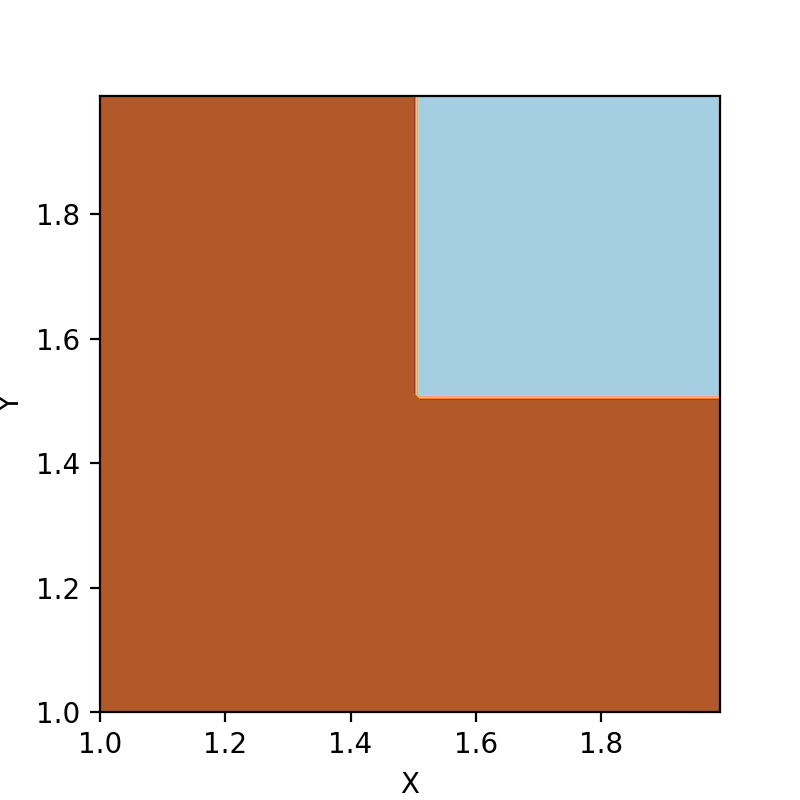


資料集三：

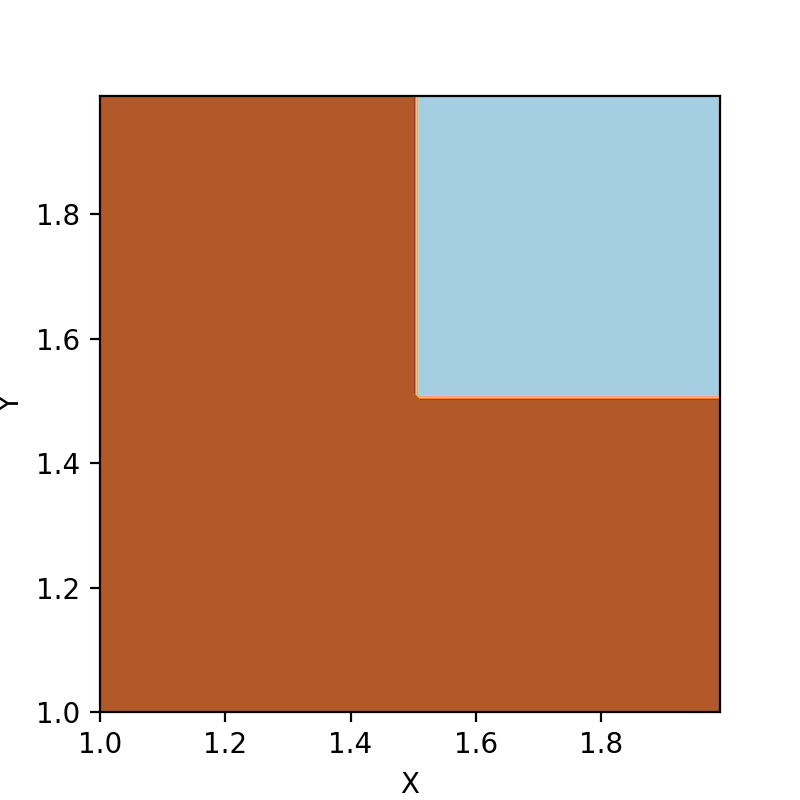
真理：如果feature1是1且feature2是2則label就是1

**資料筆數：50000 符合真理的比數佔比：3**

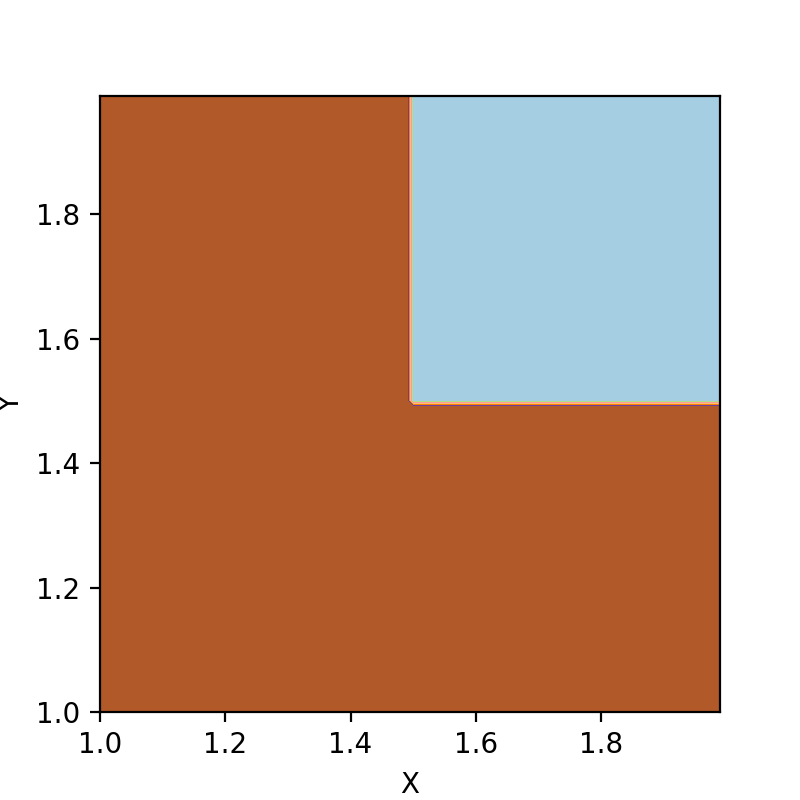
決策樹



隨機森林

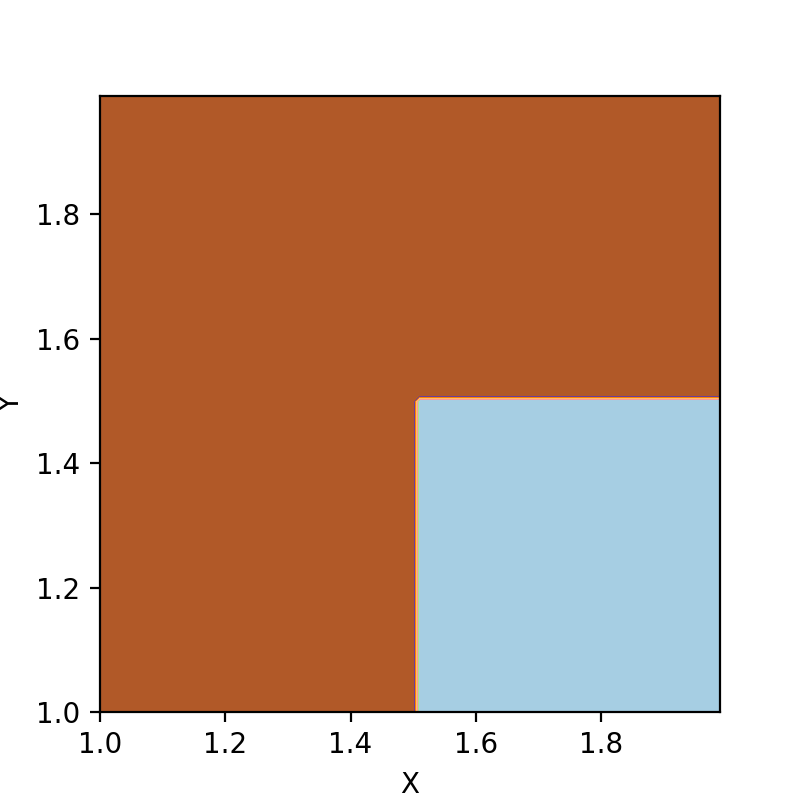


xgb

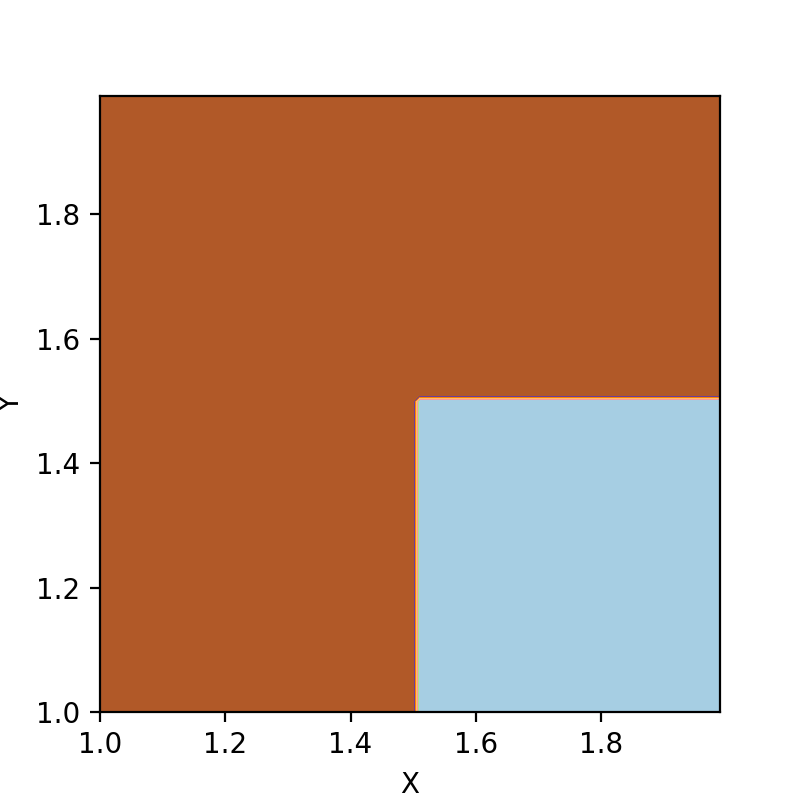


**資料筆數：50000 符合真理的比數佔比：10**

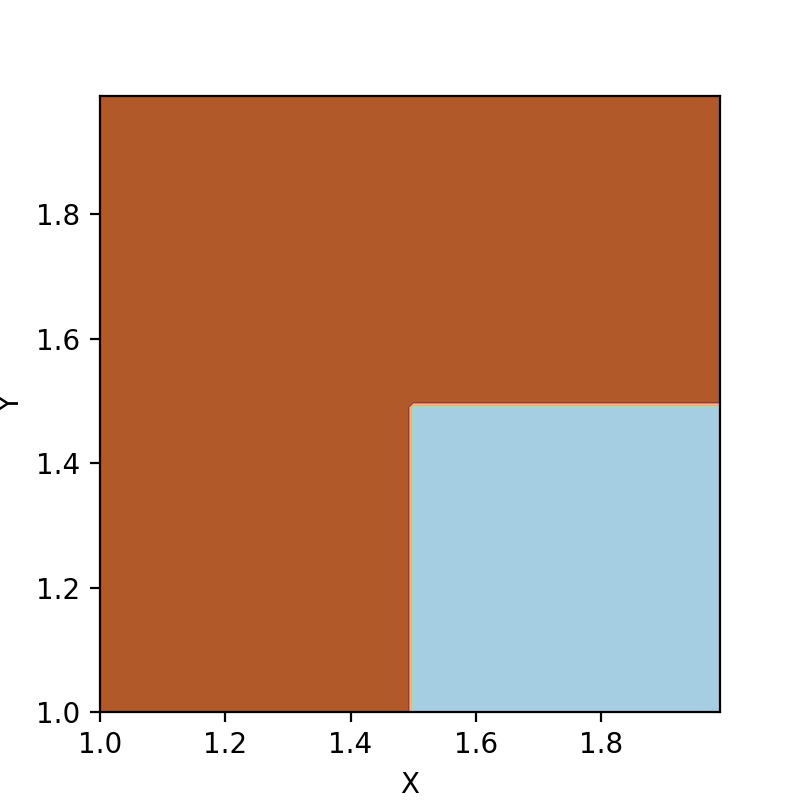
決策樹



隨機森林

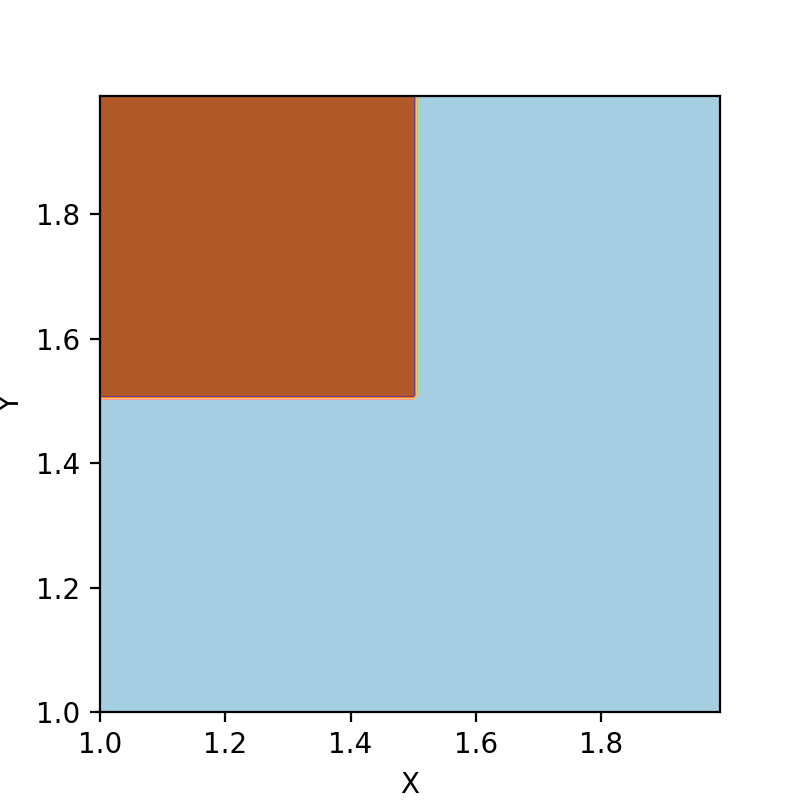


xgb

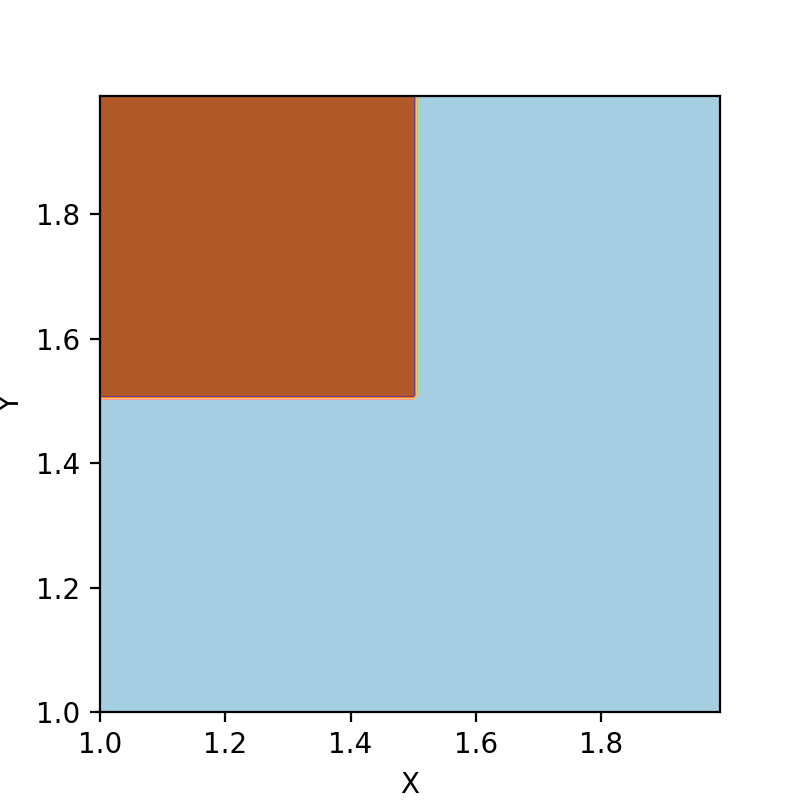


**資料筆數：50000 符合真理的比數佔比：30**

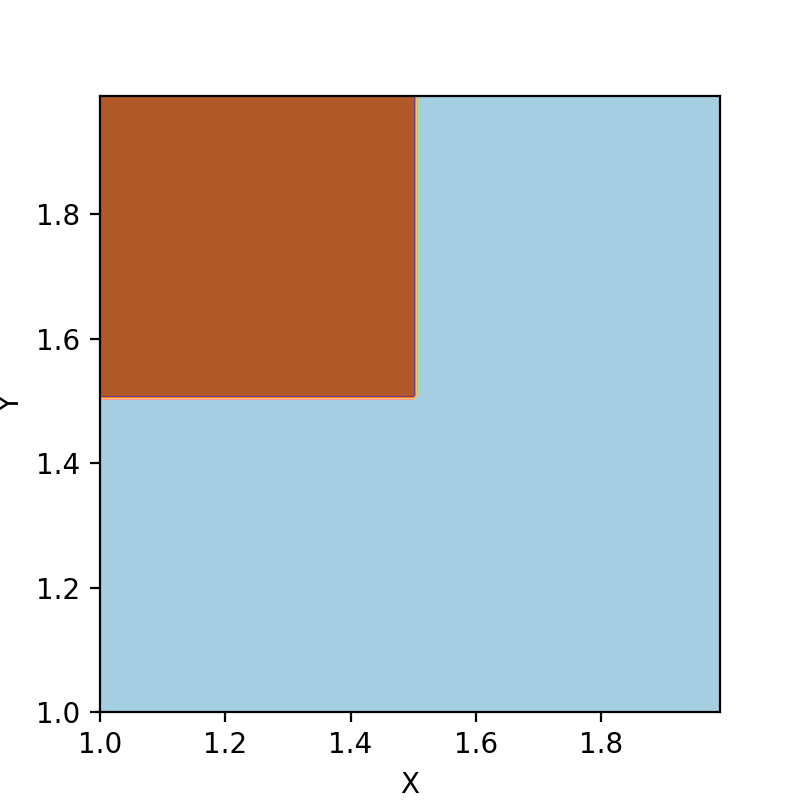
決策樹



隨機森林



xgb



結論：

1. 關於運算速度方面，決策樹最快，其次是隨機森林，最慢的則是xgboost
2. 關於三種不同資料類型的資料，再處理線性模型的時候（期中考期末考分數）在資料筆數少的時候，且真理佔比也不高的時候(0.25~0.5)，決策樹的表現略差於xgboost，感覺xgboost所學到的東西更加接近真理。而當資料量開始變多的時候，決策樹跟xgboost的效力差不多，只是隨機森林的效果越來越差，最好我給了真理佔比90倍的資料，也就是幾乎全部都是真理，只有少部分是雜訊的時候隨機森林的模型整個長歪了，他已經overfitting了，所以真理跟雜訊佔比分常重要，而xgboost在這方面抗overfitting的能力感覺較強
3. 而最後關於我自己訂的規則方面，由於資料類型非常簡單，只有0跟1所以給太多資料筆數的時候他會學到非真理的東西也當成真理，要一直到當我們把真理的數量調高之後，他才會判斷誰是真理誰是雜訊。