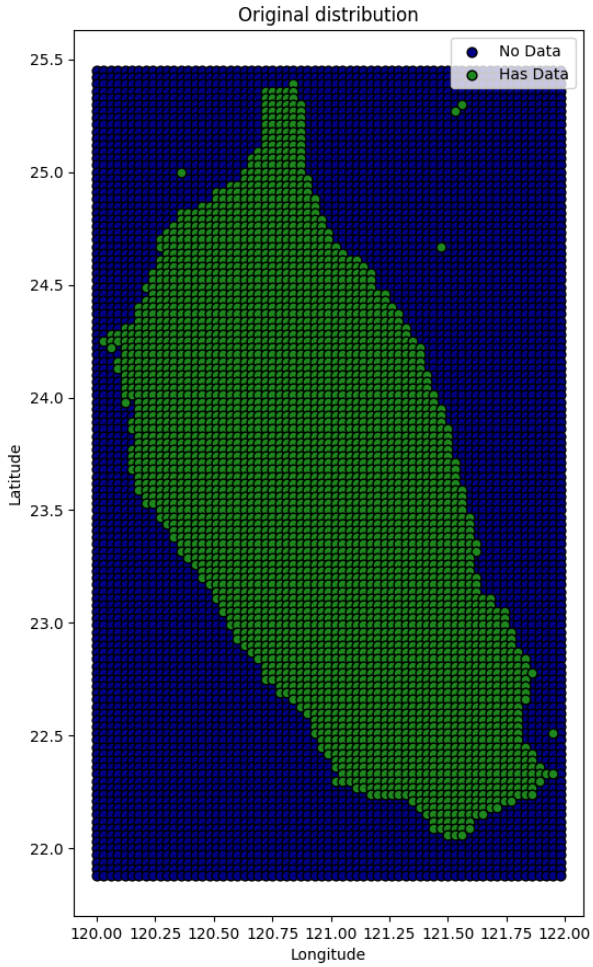
ML Assignment 4 Report

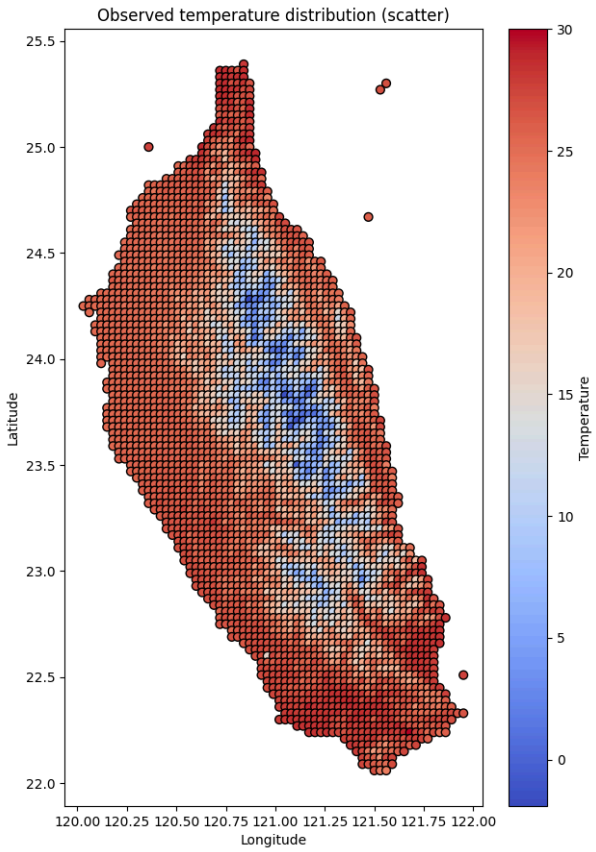
Data Preprocessing

首先，原始的XML檔內有其他標籤資訊，因此先將 `<contents>` 內的東西取出，再將分行符號換為空白，並用逗號分隔取出 \(67\times 120=8040\) 個值。
但是，每一行的最後一個值和下一行的第一個值因為中間沒有逗號，所以會連在一起，變成一個 `-999.0E+00-999.0E+00` 的字串，進而導致剛剛用逗號分隔出來的值會缺少。並且第 8040 個值沒有接續字串，所以並沒有被逗號分開，所以再手動補上。
因此這裡直接將其取代為 `-999.0E+00,-999.0E+00` ，這樣再用逗號分隔一次即可取到 8040 個元素。
接下來， `safe_float` 函數將其中的 `-999.0E+00` 值轉為 `NaN` ，再重新調整為 120 列 67 行的 DataFrame。
下一步，依照要求我們需要將資料分成兩種資料集：

- 分類資料集： `(經度, 緯度, label)`



- 迴歸資料集：（經度, 緯度, 溫度）



接下來就可以開始訓練模型。

Classification

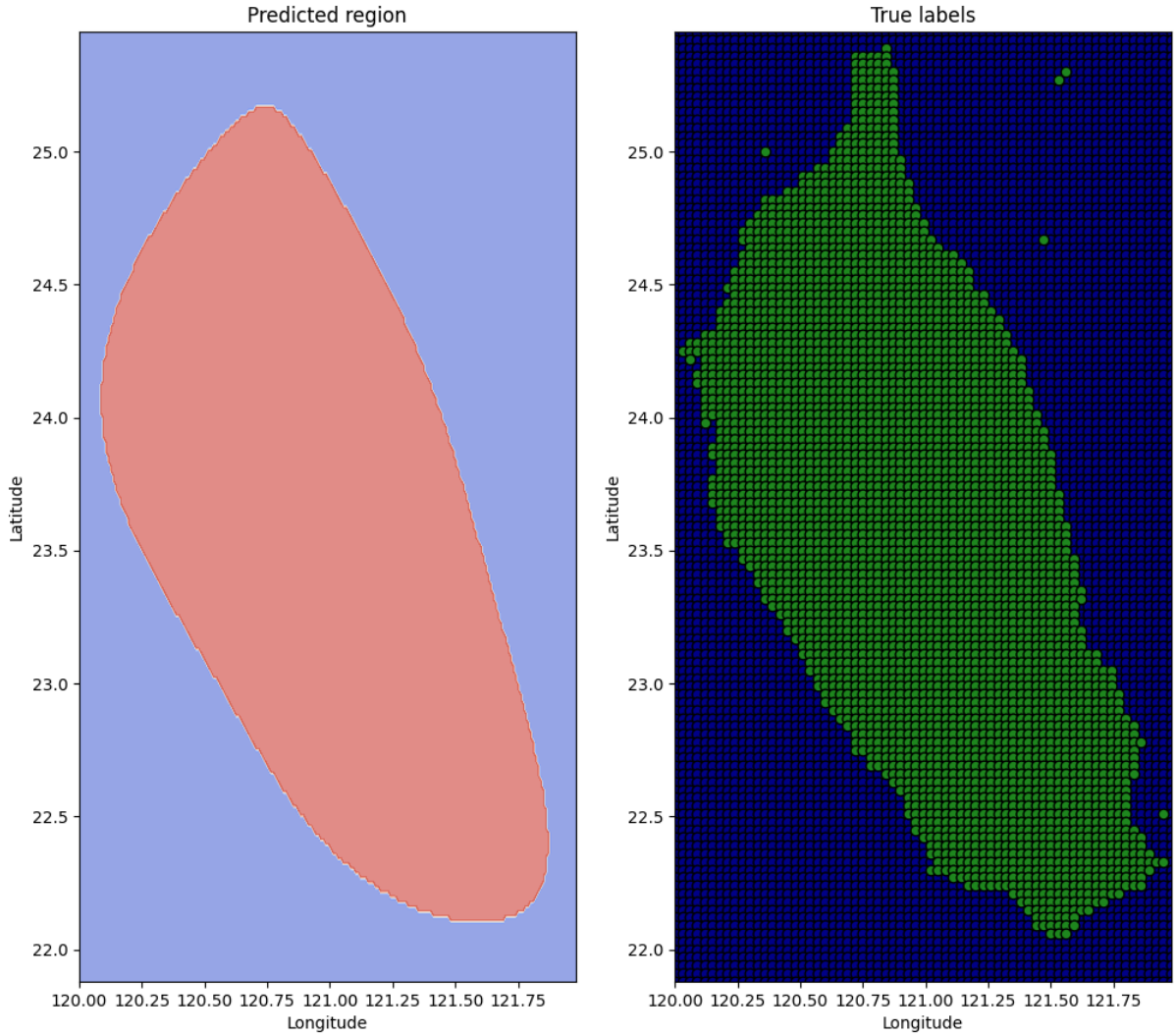
先將資料切分為80%的訓練集與20%的測試集，接下來開始訓練模型：

1. Logistic Regression：將資料送進三個程序
    - 先將經緯度的資料轉為標準化資料，讓資料的尺度小一點
    - 增加多項式次數的特徵，讓其能更好擬合原本資料的分界線
    - 讓迭代次數增加

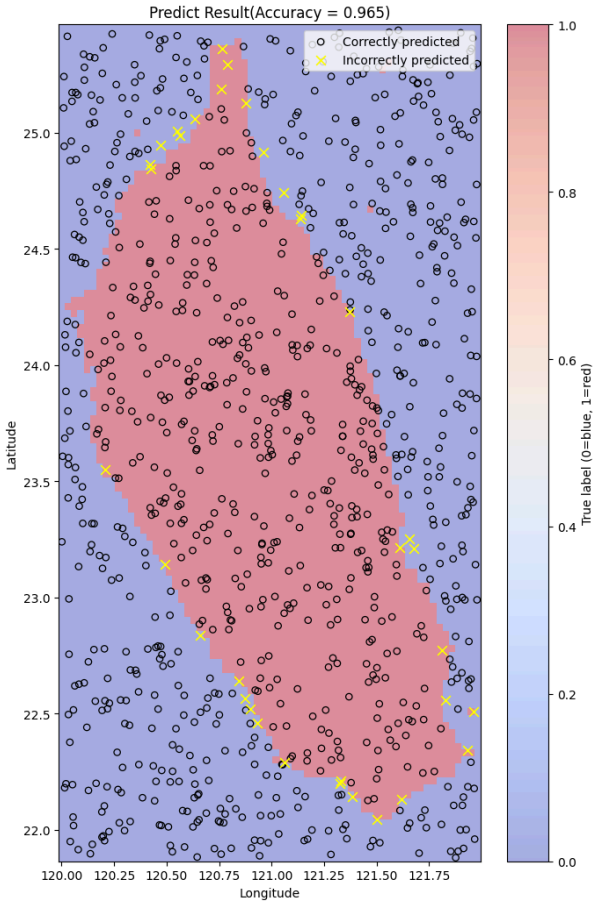
- 這裡會做上述調整是因為，直接套原本的 Logistic Regression 準確率只有略高於 50%

2. 得到訓練集準確率與測試集準確率分別為 **96.67%** 與 **95.71%**

3. 將預測區域與正確標籤一起比較



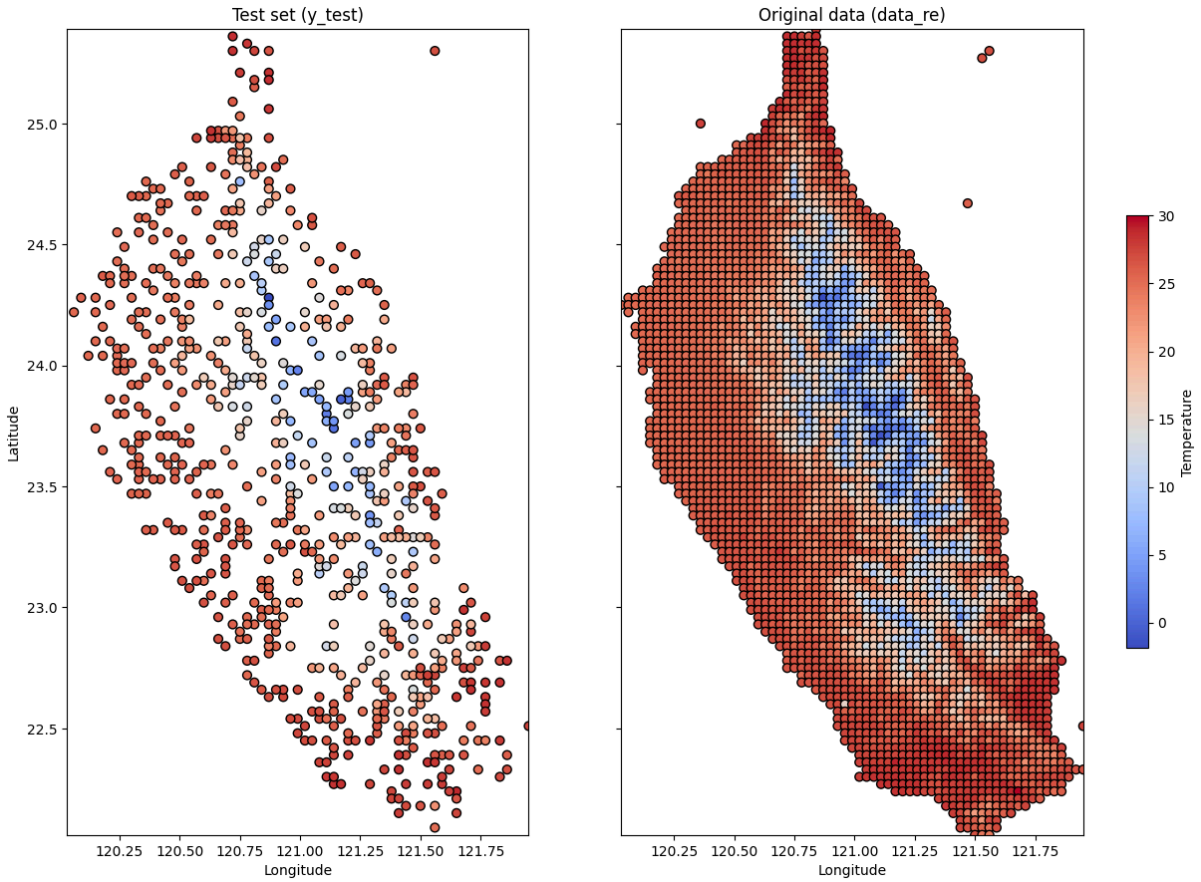
4. 隨機生成 1000 筆在此資料範圍內的經緯度並預測結果，空心圈代表預測正確，而黃色叉叉代表預測錯誤，正確率約在95%以上



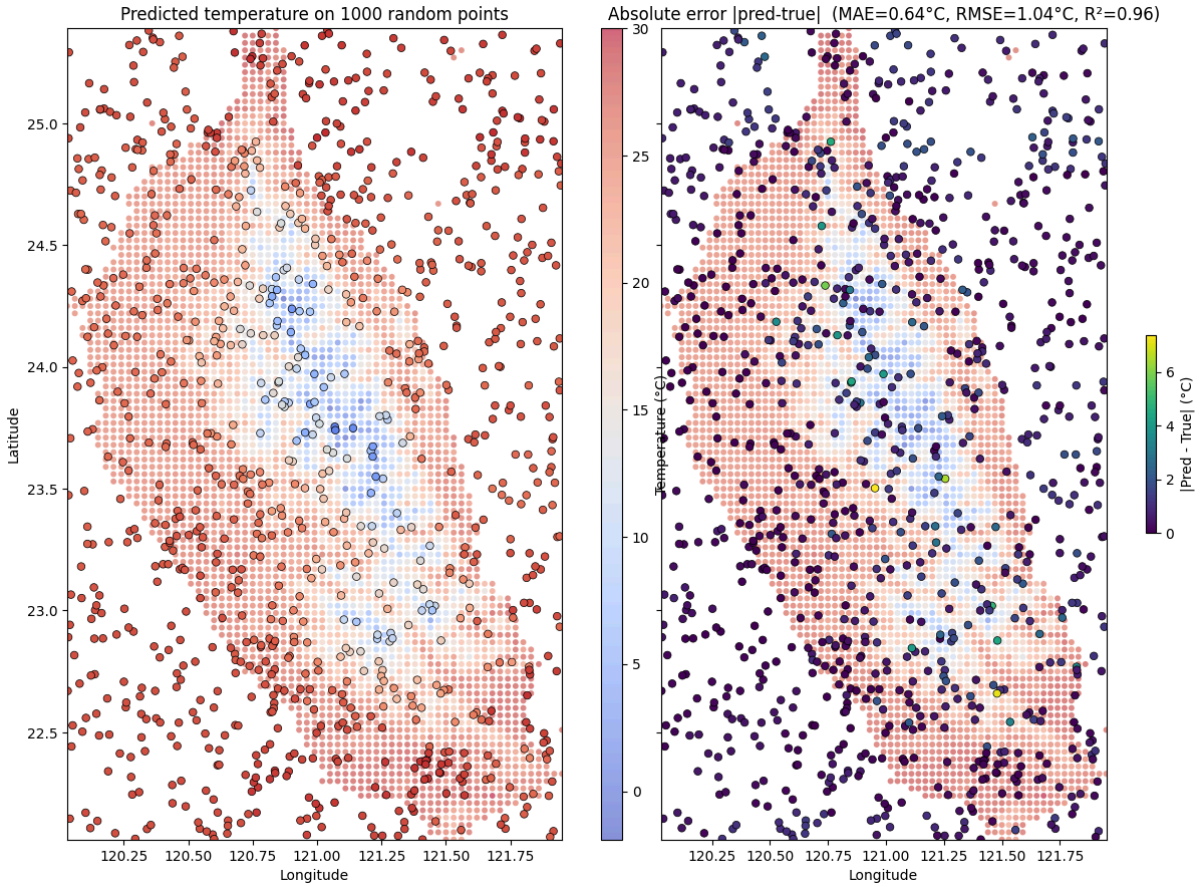
Regression

和分類時相同，將資料拆分為80%/20%，接下來開始訓練模型：

1. Random Forest Regression（這裡沒有另外調整參數）
2. 直接進行預測，得到 MSE=5.1920，R平方為0.8524。左圖為測試集預測出來的溫度，右圖為原始資料：



3. 和分類時一樣，再隨機生成 1000 個點進行預測，畫圖時背景是原始的溫度圖，左圖的實心點為預測出來的溫度，右圖的實心點為預測誤差



- `cKDTree` 是拿來找最近點用的，當然，也可以直接計算距離