氣象觀測資料分析與模型訓練報告

1 作業目標

本作業目標為處理中央氣象署的網格化溫度觀測資料。首先，對原始 XML 資料進行轉換與前處理，建立適用於機器學習的「分類」與「迴歸」兩種資料集。接著，分別使用兩組截然不同的機器學習模型——線性模型（邏輯迴歸、線性迴歸）與 K-最近鄰模型（KNN），對這兩個資料集進行訓練與評估，並深入分析與比較其效能差異。

2 資料轉換與前處理

本專案使用的原始資料為 `O-A0038-003.xml`，其中包含了台灣地區的網格化溫度觀測值。資料轉換的流程如下：

1. 資料解析：使用 Python 的 `xml.etree.ElementTree` 函式庫解析 XML 檔案，提取 `<Content>` 標籤內的網格資料字串。

2. 資料清理：原始資料字串中包含換行符號 `\n` 與逗號 `,` 作為分隔。為了正確讀取數值，程式先將所有分隔符統一處理，再將字串分割成一個包含 8040 (120×67) 個溫度值的數值列表。

3. 建立資料集：根據經緯度起始點（東經 120.00 度，北緯 21.88 度）與解析度（0.03 度），遍歷所有網格點，生成兩個 Pandas DataFrame：

    - 分類資料集 **(Classification)**：包含所有 8040 個網格點。根據作業規則，若溫度值為無效值 -999.0，則 `label` 標記為 `0`（代表無效觀測點）；反之，則標記為 `1`（代表有效觀測點）。
    - 迴歸資料集 **(Regression)**：僅保留 3495 個溫度觀測值不為 -999.0 的有效資料點，用於後續的溫度預測。

資料處理完成後，分別將兩個資料集以 80% 訓練集、20% 測試集的比例進行分割。為了更直觀地理解資料，我們繪製了資料分佈圖（見圖 1）。

3 模型說明與選擇

為了探討不同模型在地理空間資料上的表現，我們選用了兩大類基礎模型進行比較。

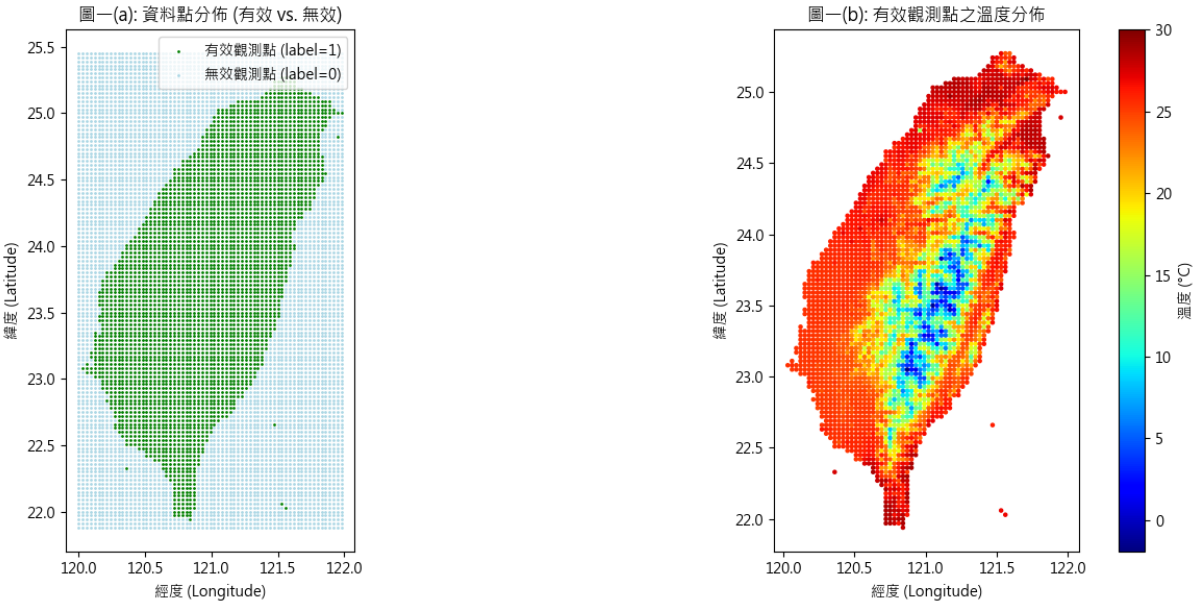


Figure 1: 資料點分佈與溫度熱力圖。圖 (a) 清晰地展示了有效觀測點（綠色）構成了台灣島的輪廓，而無效觀測點（藍色）則對應周圍海域。圖 (b) 的溫度熱力圖則直觀地顯示了溫度的地理分佈，例如中央山脈區域的溫度明顯較低（藍色區域），符合現實情況。

3.1 模型 A: 線性模型 (Linear Models)

線性模型的核心假設是特徵與目標之間存在線性關係，試圖找到一個全域適用的「最佳公式」。

- 邏輯迴歸 (**Logistic Regression**)：用於分類任務。它學習一條直線或一個平面，以將資料點劃分為兩個類別。
- 線性迴歸 (**Linear Regression**)：用於迴歸任務。它學習一個線性方程式（溫度 $= w_1 \times$ 經度 $+ w_2 \times$ 緯度 $+ b$）來預測連續的溫度值。

3.2 模型 B: K-最近鄰模型 (K-Nearest Neighbors, KNN)

KNN 是一種非參數演算法，核心思想是「物以類聚」，一個點的特性由其最鄰近的 K 個點來決定，使其能有效捕捉局部與非線性的資料模式。

- **KNeighborsClassifier**：用於分類任務，透過鄰居投票決定類別。
- **KNeighborsRegressor**：用於迴歸任務，透過鄰居數值的平均來進行預測。

4 訓練過程與結果分析

4.1 分類模型：預測資料點是否有效

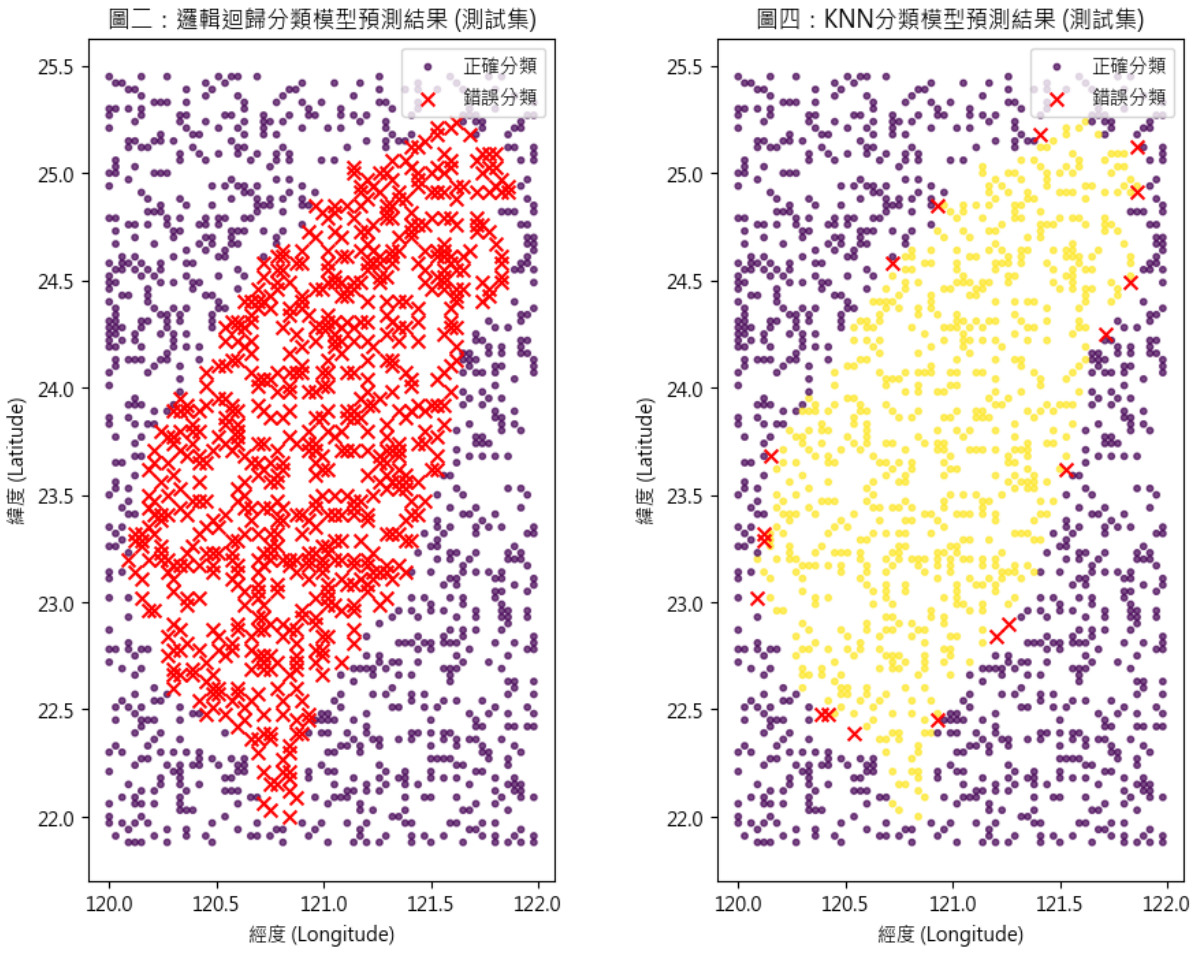
此任務的目標是根據經緯度，判斷一個網格點是否為有效觀測點。

Table 1: 分類模型評估指標比較

評估指標	**Logistic Regression**	**KNeighborsClassifier**
準確率 (Accuracy)	0.5653 (56.5%)	**0.9888 (98.9%)**
F1-Score (Label 1)	0.00	**0.99**

結果分析

- **Logistic Regression** 的表現極差：模型準確率僅 56.5%，且對於 `label=1` 的 F1-Score 為 0，表示模型完全無法辨識出任何有效資料點。圖 2a 直觀地展示了這個災難性的結果：模型幾乎將所有位於台灣輪廓內的點都錯誤地分類了（大量的紅色 X）。這是因為線性模型試圖用一條直線分割不規則的地理邊界，是根本不可能的任務。
- **KNeighborsClassifier** 的表現非常出色：準確率高達 98.9%，且各項指標均接近完美。圖 2b 顯示，絕大多數測試點都被正確分類，只有極少數的點（紅色 X）被誤判，且這些誤判點都發生在有效與無效區域的邊界地帶，證明了 KNN 模型能完美學習這種複雜的非線性邊界。



(a) 圖二：邏輯迴歸分類結果　　(b) 圖四：KNN 分類結果

Figure 2: 分類模型預測結果比較 (測試集)

4.2 迴歸模型：預測對應的溫度觀測值

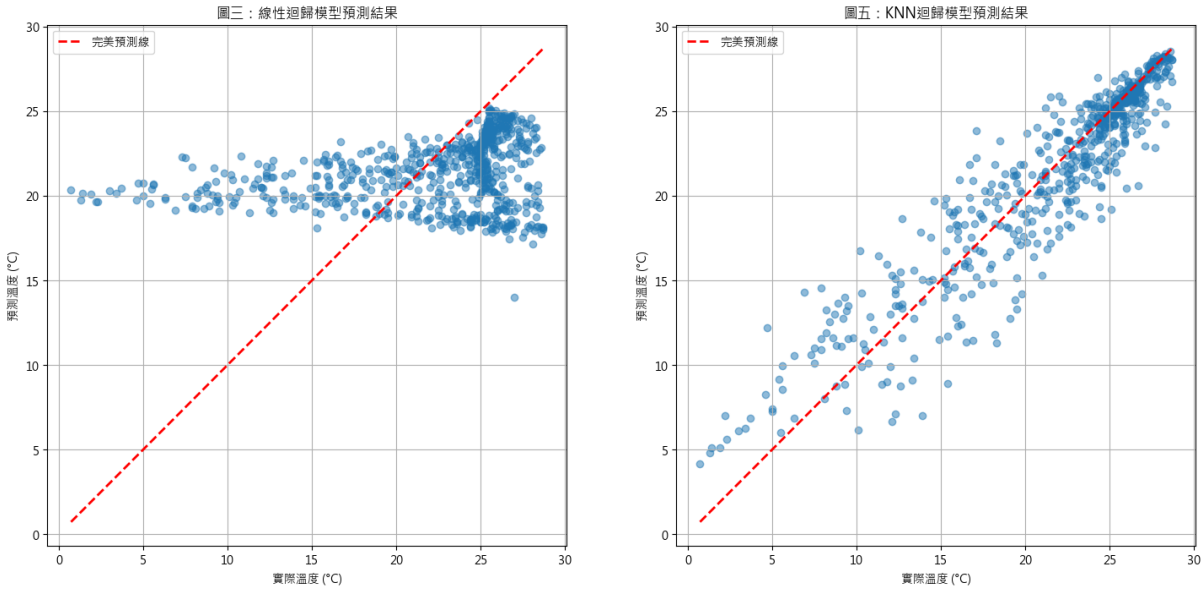
此任務的目標是根據經緯度，預測該地點的溫度。

Table 2: 迴歸模型評估指標比較

評估指標	**Linear Regression**	**KNeighborsRegressor**
均方誤差 (MSE)	32.1430	**4.4453**
R-squared (R^2) 分數	0.0525	**0.8690**

結果分析

- **Linear Regression** 的表現極差：R^2 分數僅為 0.0525，代表模型幾乎沒有預測能力。圖 3a 顯示，預測點（藍點）非常鬆散地分佈在完美預測線（紅色虛線）周圍，且整體趨勢偏離甚遠，證明了線性模型無法捕捉複雜的溫度變化。
- **KNeighborsRegressor** 的表現相當優異：R^2 分數高達 0.8690，表示模型具備很強的預測能力。圖 3b 將實際溫度與預測溫度進行比較，可以看到散點大多緊密分佈在完美預測線周圍，再次證明了 KNN 模型的優異性能，能夠有效預測局部溫度。



(a) 圖三：線性迴歸預測結果　　(b) 圖五：KNN 迴歸預測結果

Figure 3: 迴歸模型預測結果比較 (實際 vs. 預測)

5 結論

本專案成功完成了資料轉換與模型訓練的任務，並透過視覺化圖表得到一個非常清晰的結論：對於具有複雜、非線性模式的地理空間資料，非線性的 **K-**最近鄰 **(KNN)** 模型效能遠優於基礎的線性模型。

- 在分類任務中，KNN 能精準學習由資料點構成的不規則邊界，而邏輯迴歸則完全失效。
- 在迴歸任務中，KNN 能有效捕捉局部的溫度變化，具備高度的預測能力，而線性迴歸無法處理非線性的溫度分佈，其預測結果不具參考價值。

這個結果突顯了在應用機器學習時，「選擇適合資料特性的模型」是至關重要的第一步。對於此類問題，應優先考慮能夠處理非線性與局部特性的模型。