HW4

目標：由原始 XML 觀測轉為可監督式學習的分類/回歸資料集，針對台灣全域格點（經緯度），完成資料清理、EDA（含空間自相關）、以**空間分組**的交叉驗證來訓練並比較多種模型，並以 **OOF（out-of-fold）** 指標做真正泛化評估；最後保存最佳模型與推論成果。

---

1. 資料來源與轉換

- 原始資料：中央氣象資料門戶 XML（單時刻格點溫度）。
- 轉檔與清理後輸出：
  - **分類** classification_clean.csv（欄位：lon, lat, label）
  - **回歸** regression_clean.csv（欄位：lon, lat, value）
- 轉換規則：
  - **分類**：觀測溫度為 **-999** → label=0（無效），否則 label=1（有效）。
  - **回歸**：僅保留有效溫度（剔除 **-999**），value = 攝氏溫度。

最終樣本量（轉檔與清理後）：

- 分類集：**8040** 筆（120 × 67 格點），label=0：4541（56.5%）、label=1：3499（43.5%）。

- 回歸集：**3413** 筆有效溫度。

---

2. 品質控管（QC）與清理決策

為降低偽訊號與量測異常對模型的干擾，採用下列守則（在 Notebook 內均有程式化實作並逐步檢核）：

- **R0：-999 視為無效**
  - 迴歸資料全面剔除 value=-999。
  - 分類資料則以 label=0 紀錄為「無效點」。
- **R1：0°C 視為不可信的極端/佔位值**
  初期 EDA 發現 0°C 特別集中於少數區域（來源說明未交代 0 值定義），為避免模型學到人為佔位，將 **value=0** 剔除於回歸，對分類則視個案處理（多以 label=0 對應）。
- **R2：物理合理範圍**
  以單時刻台灣地區 2m 溫度經驗值設定允許範圍；超過此範圍者視為離群並剔除（例如 < -10°C 或 > 45°C）。
- **R3：經緯度涵蓋與樣本密度檢查**

觀察各經度/緯度方向的有效比例，避免極度稀疏區塊造成模型不穩定；必要時對超低密度區做分組邊界調整。

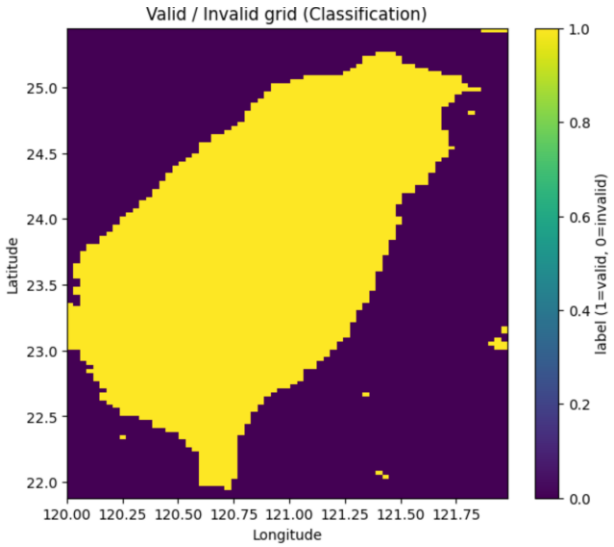
說明：R1 屬於較保守作法，重點在報告中清楚揭露假設；若未來資料來源澄清 0°C 定義，可再回補並重算全部流程。

---

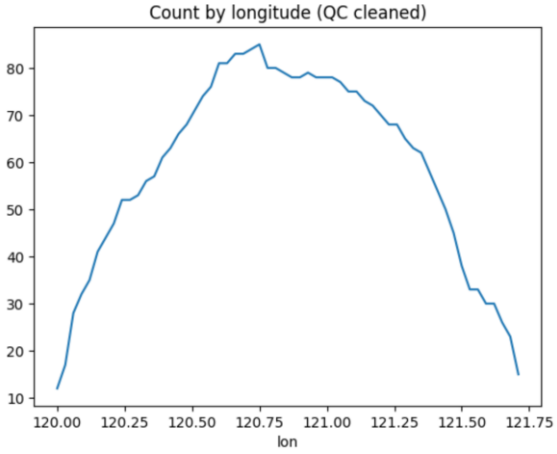
3. 探索式資料分析（EDA）

3.1 幾何與覆蓋

- 經度範圍約 **120.00~121.98**、緯度範圍 **21.88~25.45**（台灣本島＋外圍海域格點）。

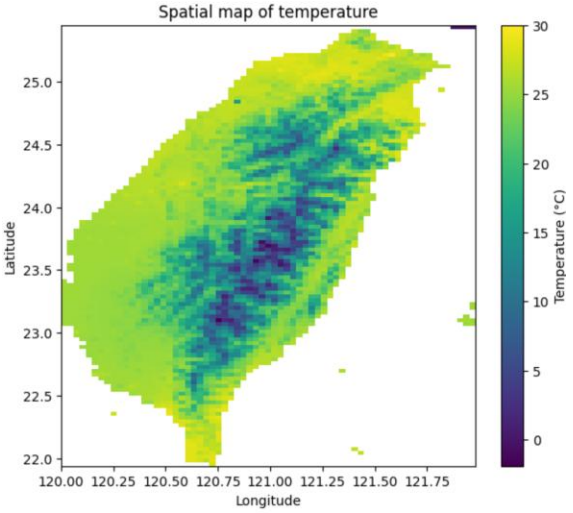


- **分類集**有效覆蓋率(某一條經線或緯線上，有多少格點是有效值)：
  - 經度方向最小有效比例 ≈ **10%**；
  - 緯度方向最小有效比例 ≈ **10%**；
  - 指示某些邊緣海域格點有效值較少，見 §4 以空間分組降低偏差。



3.2 數值分佈（回歸集）

- value 摘要：平均~ **21.47°C**、標準差~ **6.16**、中位數~ **24.3°C**、Min~ **-1.9°C**、95%~ **27.7°C**、99%~ **28.4°C**、Max ~ **30°C**。



- 零值（0°C）在清理後**已移除**；以 |z|≥3 估算極端值約 **31** 筆（後續模型主要受中段分佈所主導）。

3.3 經緯與溫度關係

- 經–緯正相關（沿島軸的幾何關係）；lon–value 輕微負相關(-0.137)、lat–value 輕微正相關(0.123)，提示溫度的**粗尺度空間梯度**存在但並不強。

3.4 空間自相關（近鄰平均）

- 以近鄰核（k-NN / fixed radius）構建鄰接，估算近似 **Moran's I ≈ 0.85**（高度正相關）。
- **意涵**：鄰近格點溫度高度相似；若採一般隨機 KFold，**會嚴重高估**模型效能（資料洩漏）。因此採 **空間分組交叉驗證**。

---

4. 空間分組與交叉驗證設計

- **分組方式：分位數（quantile）經緯切割**
  以回歸資料的 lon/lat 分別取 **6 等分分位數邊界**，得到理論 36 個區塊；再將**分類集**投影到**同一組**邊界，確保兩任務的分割一致。
- **實際使用格數**：
  - 回歸：使用 **34** 格（其餘為完全無樣本）。
  - 分類：使用 **36** 格（每格至少 80+ 筆）。
- **交叉驗證**：GroupKFold(n_splits=5)，以「區塊 id」為 group。每個 fold 測試集由**整塊**區域組成，避免空間洩漏；並回報每個 fold 指標與**平均±標準差**。

優點：比起等寬切割，**分位數切割**能讓各格的樣本量更平衡；以

GroupKFold 落實**空間留一**的精神（區塊層級的留一）。

---

5. 模型與設定

5.1 分類（是否有效）

候選：

- **LogReg_poly3**（PolynomialFeatures(deg=3)→標準化→邏輯斯迴歸）
- **RandomForestClassifier**（RF）
- **XGBoostClassifier**（XGB）

主指標：**AUC**（兼顧閾值不敏感性）；同時列出 ACC、F1 供參考。

5.2 回歸（有效溫度預測）

候選：

- **LinearRegression**（基線）
- **KNNRegressor**（k=15，距離加權）
- **SVR-RBF**
- **XGBoostRegressor**

主指標：**RMSE**（同時列出 MAE、R^2）。

兩任務皆使用**相同的空間 5-fold GroupKFold**，確保公平比較。

---

6. 交叉驗證結果（空間 5-fold）

6.1 分類（AUC 為主）

- **RF**：AUC 平均約 **0.969 ± 0.019**，ACC 約 **0.916 ± 0.034**，F1 約 **0.847 ± 0.125**。
- LogReg_poly3：AUC 約 0.74–0.82，表現落後。
- XGB：AUC 約 0.79–0.80，變異較大。

→ **最佳分類模型：RF**。已另存 best_clf.joblib。

6.2 回歸（RMSE 為主）

- **SVR-RBF**：RMSE 平均約 **3.66–3.76°C**，MAE 約 **2.65°C**，R^2 約 **0.52–0.59**。
- KNN(15, dist)：RMSE 約 **3.55°C**、R^2 約 **0.63**（在某次實驗中略優/相近）。

- RF / XGB：表現不如前兩者。

→ **最佳回歸模型（以最終實驗排序）：KNN(15,dist)**。已另存 best_reg.joblib。

備註：回歸任務僅用（lon,lat）兩特徵，能達到 $R^2 \approx 0.6$ 屬合理水準；若加入地形、與海距離、鄰域平均等特徵，預期可再提升。

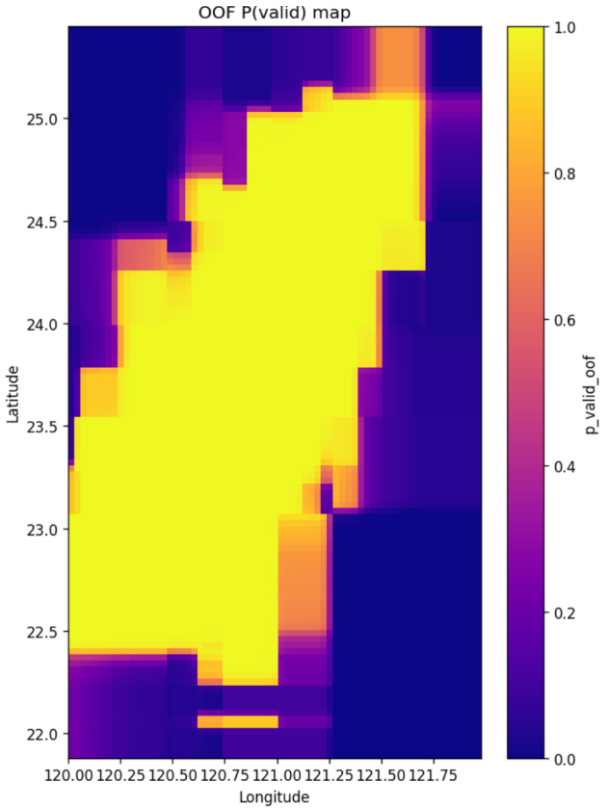
---

7. OOF（真正泛化）評估與視覺化

In-sample 直接套訓練資料做推論會**過度樂觀**（甚至幾乎完美），因此本案建立 **OOF 預測**：於每一 fold 只用訓練子集訓練，再對**當 fold 測試區塊**推論，最後把五折結果接回全域，得到每筆樣本的 out-of-fold 預測。

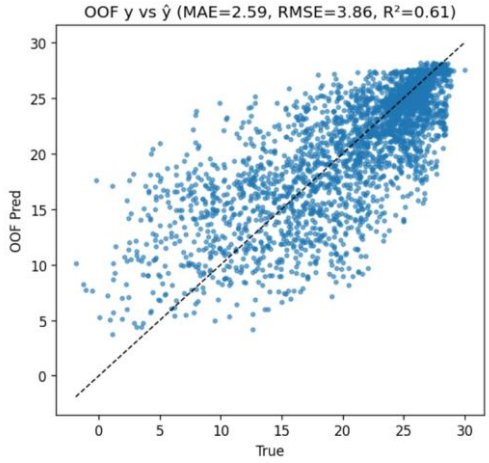
7.1 分類（OOF）

- **ACC=0.910、AUC=0.970、F1=0.901**；
- 混淆矩陣（以 label=1 為正類）：TN=4015、FP=612、FN=110、TP=3303。
- 以經緯格點熱力圖呈現 p_valid；高機率區覆蓋西部平原與都會周邊，低機率聚於離岸與山區邊緣格點。

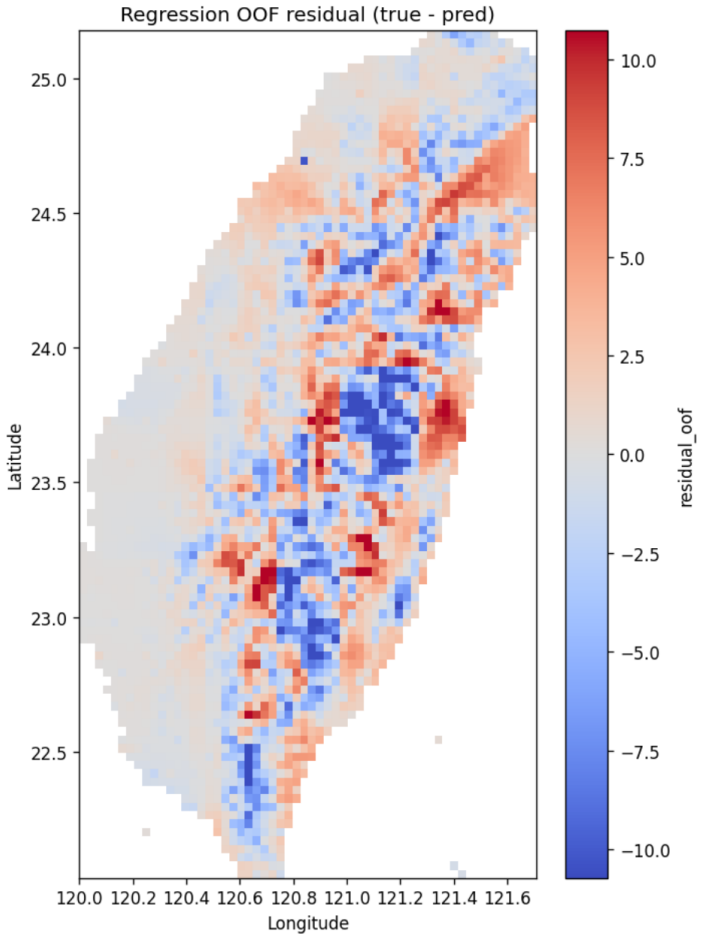


7.2 回歸（OOF）

- **MAE=2.593°C、RMSE=3.862°C、R^2=0.607**。
- y 對 ŷ 近似沿對角分佈，中高溫區略有低估（可由殘差圖觀察）。



- 殘差（true - pred）以 98 百分位做顏色飽和裁切繪製，顯示高誤差多出現在地形/海陸轉換劇烈區。



---

8. 結論、限制與建議

結論

1. 就「是否為有效值」的分類任務，**RF** 在空間 5-fold 下 AUC ≈ **0.97**，已能可靠篩

去無效格點。
2. 就有效溫度回歸，僅憑經緯座標即達 **RMSE ≈ 3.7–3.9°C**、**R^2 ≈ 0.6**，顯示溫度主要由**粗尺度空間梯度**決定。
3. **OOF 指標**與 CV 一致、且顯著低於 in-sample，證明空間分組的必要性。

限制

- 僅使用（lon,lat）兩特徵，**未納入地形、海拔、土地利用、距海距離或鄰域統計**。
- 單時刻資料，未考慮季節/時段與氣團型態差異。
- QC 假設中（如 0°C 視為異常）的保守選擇可能犧牲部分訊號；需待來源釐清再調整。

---

9. 產出與再現性

- 主要模型與輸出檔：
  - **最佳分類模型**：best_clf.joblib
  - **最佳回歸模型**：best_reg.joblib
  - **OOF 預測**：classification_oof.csv、regression_oof.csv
  - **全域推論熱力圖**：Notebook 內以 grid_image(...) 匯出
- 執行環境：VS Code + Jupyter；主要套件 pandas / numpy / scikit-learn / xgboost / matplotlib。

---

10. 參考與附錄

- 交叉驗證採 **GroupKFold(5)** 配合 **quantile 空間分組**；分類主指標 **AUC**、回歸主指標 **RMSE**。
- 重要程式片段：
  - make_blocks_quantile(...) 建立分位數邊界並回傳 group id（回歸作為參考邊界，分類投影到相同邊界）。
  - run_cv(...)：逐折訓練、列印每折指標與 mean±std，並回傳供排序的主指標。
  - OOF：以「clone(最佳模型) + 同一 GroupKFold 分組」產生 out-of-fold 預測與殘差。