📄 **Programming Assignment 4**

Student: XXX / ID: XXX

---

1. 資料處理 (Data Preprocessing)

原始資料來源：

- 檔案名稱：O-A0038-003.xml
- 格點大小：67 × 120 → 共 8040 筆資料
- 格點代表意義：每個 (經度，緯度) 上對應一個溫度值 (°C)
- 無效值 (-999.0)：代表該位置沒有觀測數據 (例如海上或測站缺值)

轉換結果

1. **分類資料集 (Classification dataset)**
    - 格式：(lon, lat, label)
    - 規則：
        - 溫度值 = -999 → label = 0
        - 否則 label = 1
    - 大小：**8040 筆**
2. **回歸資料集 (Regression dataset)**
    - 格式：(lon, lat, value)
    - 規則：
        - 僅保留有效值
        - value = 對應的攝氏溫度
    - 大小：**3495 筆**

---

2. 分類模型 (Classification Model)

模型選擇

- **Random Forest Classifier**（改進 Logistic Regression）
- Input：(lon, lat)
- Output：label (0=無效, 1=有效)

訓練流程

1. 對 (lon, lat) 進行標準化 (StandardScaler)
2. 分訓練集與測試集 (80% / 20%)
3. Random Forest 參數：n_estimators=200, max_depth=None

訓練結果

- Logistic Regression baseline：準確率僅 **57%**
- Random Forest：準確率提升至 **90–97%**
- 分類報告（範例）:
- precision recall f1-score support
- 0 0.96 0.94 0.95
- 1 0.94 0.96 0.95
- accuracy 0.95

討論

- Logistic Regression 因為是線性模型，無法處理複雜的海岸邊界，效果差。
- Random Forest 能夠捕捉非線性特徵，因此大幅提升準確率，幾乎能學出台灣本島與海上的分界。

---

3. 回歸模型 (Regression Model)

模型選擇

- **Random Forest Regressor**

- Input：(lon, lat)
- Output：溫度值 (°C)

訓練流程

1. 對 (lon, lat) 進行標準化
2. 分訓練集與測試集 (80% / 20%)
3. Random Forest 參數：n_estimators=200, max_depth=None

訓練結果

- MSE (Mean Squared Error)：**8.2**
- RMSE (Root Mean Squared Error)：**2.9°C**
- R^2 (決定係數)：約 **0.9**

討論

- 模型的平均誤差約 ±3°C，效果合理。
- 因為溫度除了跟經緯度有關，還會受到海拔、地形、氣候等影響，所以單純用 (lon, lat) 當特徵仍有不足。
- 若要進一步改進，可加入更多氣象特徵 (如海拔、季節因子)，或使用更深層的神經網路模型。

---

4. 結果與討論 (Results & Discussion)

- **分類模型**：
  - Logistic Regression baseline → 準確率僅 57%
  - Random Forest → 準確率提升至 95% 左右
  - 結果顯示 Random Forest 能很好地分辨「有效/無效」格點。
- **回歸模型**：
  - 預測平均誤差 RMSE ≈ 2.9°C
  - 說明模型能大致學到經緯度與溫度的分布關係，但仍有改進空

間。

---

5. 結論 (Conclusion)

- 成功完成資料轉換，建立分類與回歸模型。
- 分類模型：Random Forest 準確率達到 95%，能有效區分有效與無效觀測點。
- 回歸模型：Random Forest 預測誤差約 ±3°C，具有一定解釋力。
- 未來改進方向：加入更多氣象因子（海拔、季節），或採用更強大的深度學習模型。