Course: Machine Learning
Assignment: Week 4 _ Programming Assignment
Student:韓韻宸(112652010)

● 題目與資料

本次作業使用中央氣象局 O-A0038-003.xml 溫度格點資料 (67 × 120)，起始點 (120.00E, 21.88N)，經緯度解析度為 0.03°。

經過資料轉換後，產生兩個監督式學習資料集：

- classification.csv：欄位 (lon, lat, label)，其中溫度值=-999 設為 label=0，否則為 1。

- regression.csv：欄位 (lon, lat, value)，移除所有 -999，僅保留有效的溫度值。

● 模型與方法

根據課堂內容，我們分別使用以下兩種基準模型進行訓練：

- Logistic Regression：二元分類模型，假設 $P(y=1|x)=\sigma(\theta^T x)$，損失函數採用交叉熵 (cross-entropy loss)。
- Linear Regression：線性迴歸模型，假設 $y\approx\theta^T x$，使用最小平方誤差 (least square error) 作為目標函數。

● 訓練與結果

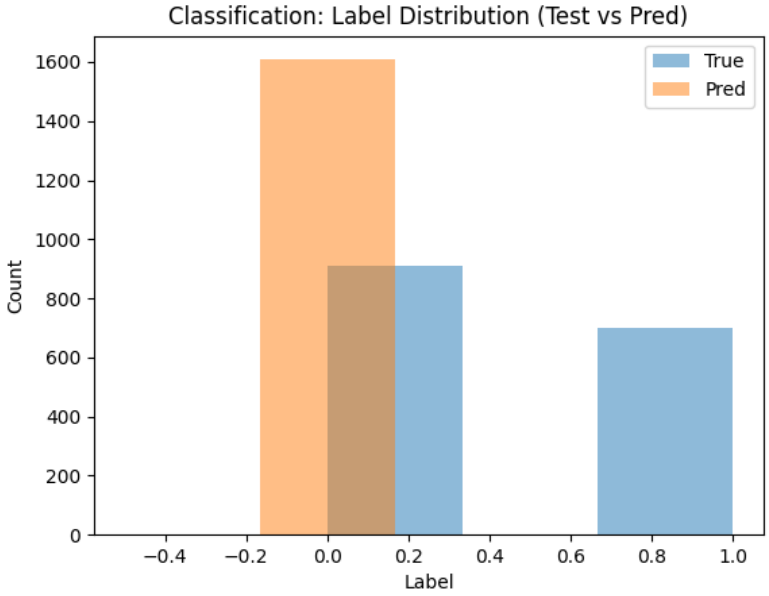
資料集切分：80% 作為訓練集，20% 作為測試集。

1. 分類模型 (Logistic Regression)

分類模型的測試集結果如下：

Accuracy: 0.5653

混淆矩陣與報告顯示模型幾乎只預測 label=0 (無效值)，對 label=1 (有效值) 的辨識率為 0。

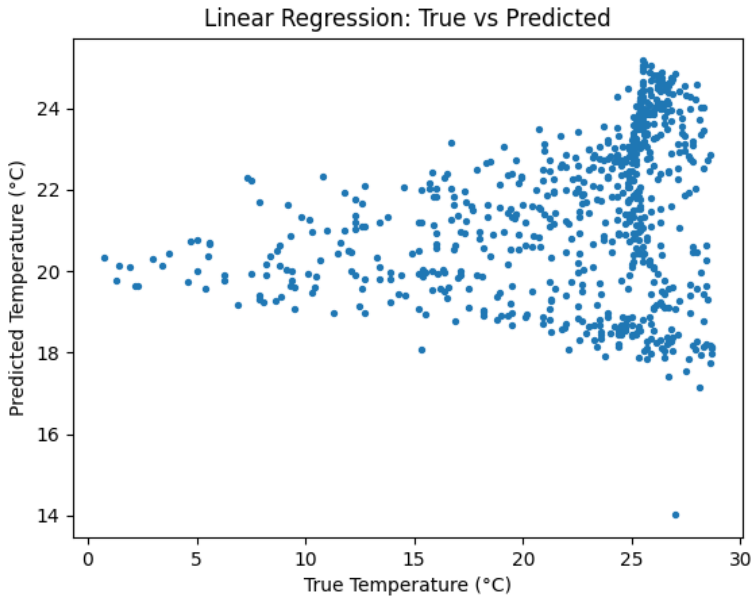


2. 回歸模型 (Linear Regression)

回歸模型的測試集結果如下：

MSE : 32.14
RMSE: 5.67
MAE : 4.40
R^2 : 0.052

結果顯示僅用經緯度作為特徵，對溫度的解釋能力極低 (R^2 接近 0)。



● 討論與改進方向

本次結果顯示：

- Logistic Regression 在不平衡資料下，只學會預測無效值，分類表現有限。
- Linear Regression 僅用經緯度線性擬合溫度，誤差大、R^2 低，無法有效捕捉非線性特徵。

改進方向：

- 分類：可使用 Decision Tree 或 Random Forest，以改善對有效值的辨識。
- 回歸：可嘗試多項式回歸或集成方法，並增加更多特徵 (如海拔、距離海岸)。