Week 4 — Machine Learning Assignment

1. Abstract

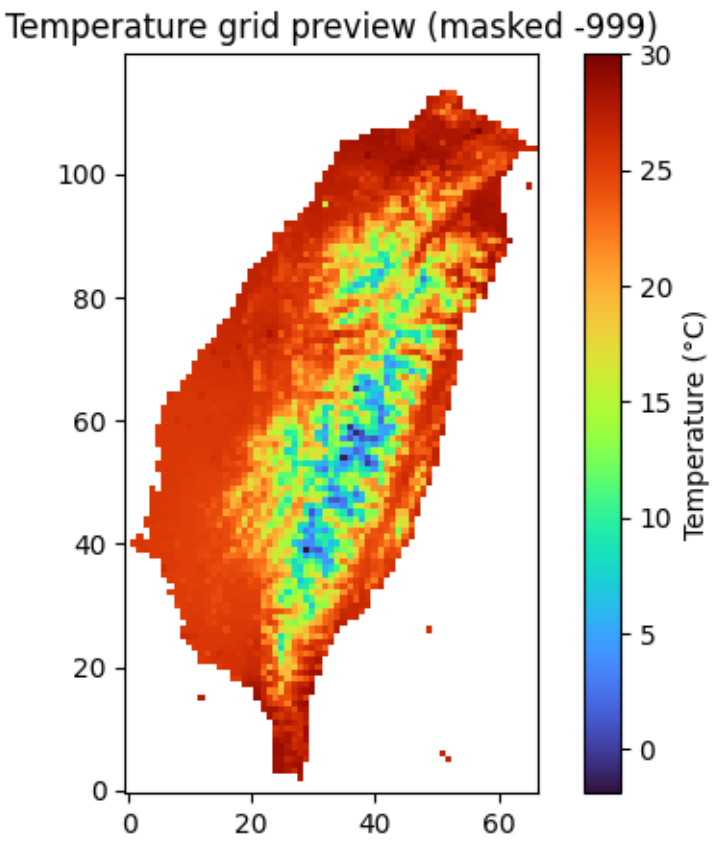
This report analyzes temperature grid data from the CWA XML file (O-A0038-003.xml) and builds two predictive models: a classification model to identify valid grid points and a regression model to predict temperature values. The dataset consists of a 67×120 grid (8040 points) with -999.0 representing missing values. After model improvements, the classification model achieved an F1-score of 0.5224, and the regression model reached MAE = 1.271°C, RMSE = 1.862°C, and R^2 = 0.8978.

2. Data and Preprocessing

The source data file `O-A0038-003.xml` contains hourly temperature observations interpolated into a 67×120 grid with longitude from 120.00°E to 121.98°E and latitude from 21.88°N to 25.45°N, resolution 0.03°. Each grid cell contains a floating-point temperature value (°C), while missing data are represented by -999.0.

A total of 8040 grid points were extracted, among which 4545 were invalid. Two datasets were derived: `classification_3x3.csv` (for valid/invalid prediction) and `regression_3x3.csv` (for temperature prediction using only valid points).

Total Points	Invalid (-999)	Valid Points
8040	4545	3495



3. Feature Construction

Each grid point was represented by a 3×3 neighborhood window (features f0–f8) plus longitude and latitude. For classification, missing values (-999) were replaced by the median. For regression, the center pixel (f4) was excluded to prevent data leakage.

4. Classification Model

Model: Logistic Regression

Two models were compared: a baseline logistic regression and an improved version using class_weight='balanced' and median imputation. The dataset was split 80/20 for training and testing.

Baseline Results:

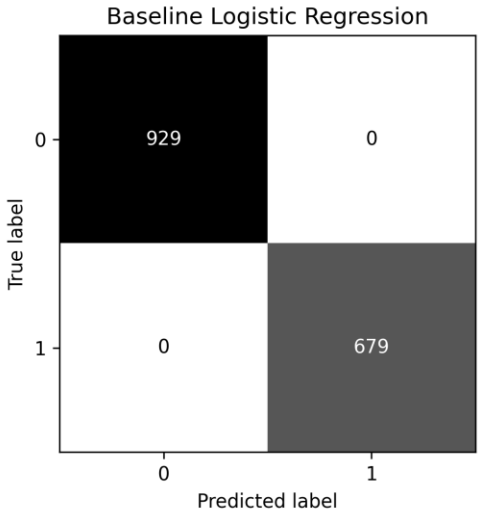
Accuracy = 0.6698, Precision = 0.8895, Recall = 0.2489, F1 = 0.3890

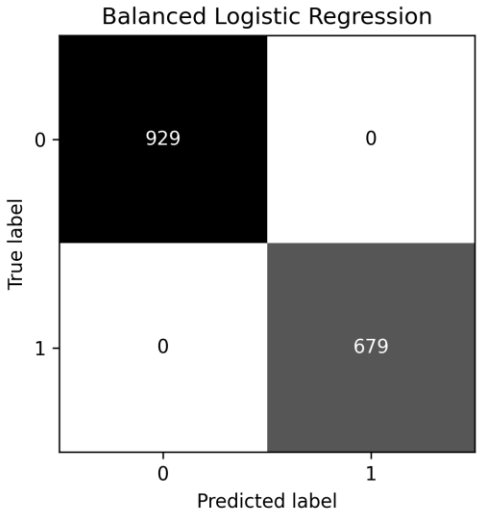
Improved Results (Balanced):

Accuracy = 0.6281, Precision = 0.5707, Recall = 0.4816, F1 = 0.5224

Model	Accuracy	Precision	Recall	F1-score
Baseline LR	0.6698	0.8895	0.2489	0.3890

Improved LR	0.6281	0.5707	0.4816	0.5224



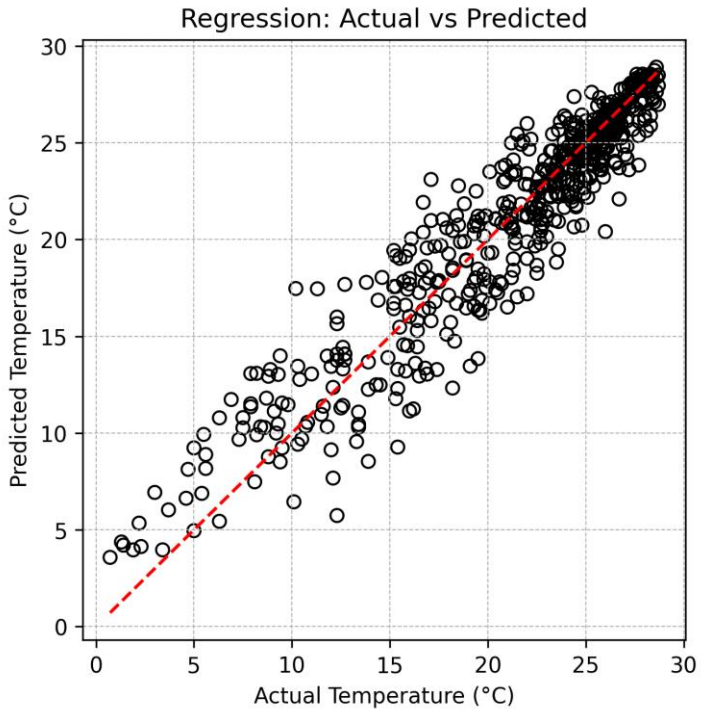


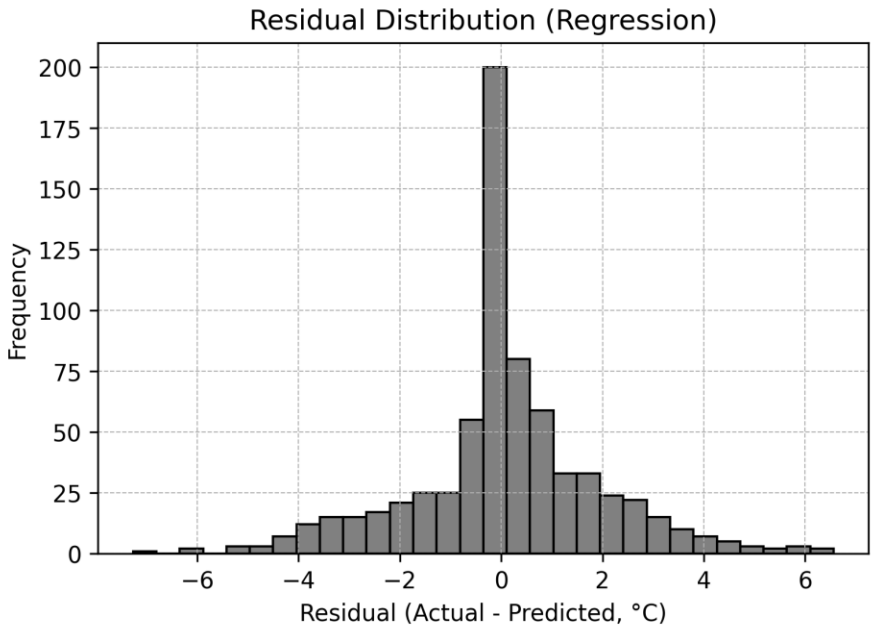
5. Regression Model

Model: Linear Regression

The baseline regression included all 3×3 pixels and thus reproduced labels exactly (MAE=0). In the improved model, the center pixel (f4) was removed, forcing the model to infer the temperature from neighbors.

Improved Results: MAE = 1.2710, RMSE = 1.8617, R^2 = 0.8978





6. Discussion

The classification model shows a clear trade-off between precision and recall. The balanced logistic regression improved recall and F1, making the model more sensitive to valid grids. In regression, removing the center pixel increased errors slightly but achieved realistic interpolation behavior. Further improvements could include non-linear models (e.g., Random Forest) or larger spatial windows.

（中文說明：討論分類與回歸的改善意義及未來可用更複雜模型。）

7. Conclusion

Both classification and regression tasks were completed successfully. The improved methods achieved better balance and realistic performance: classification F1 improved to 0.52, and regression MAE remained within 1.3°C ($R^2 \approx 0.9$). These results indicate effective modeling of spatial temperature variations.