HW3_Report

110652012 施品光

1. Introduction

In this assignment, we worked with a meteorological XML dataset containing temperature grid values over Taiwan. The grid size is 67×12067 \times 12067×120, and each grid point is associated with a longitude, latitude, and a recorded temperature value. However, many grid points contain the invalid value -999.0, which indicates missing data.

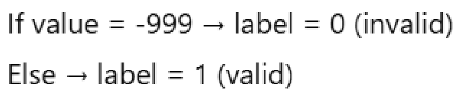
The task is divided into two supervised learning problems:

1. **Classification**: Determine whether a grid point is valid (label = 1) or invalid (label = 0).
2. **Regression**: Predict the actual temperature value at valid grid points.

2. Data Preprocessing

2.1 Classification Dataset

- Format: (lon, lat, label)
- Rule:



If value = -999 → label = 0 (invalid)
Else → label = 1 (valid)

2.2 Regression Dataset

- Format: (lon, lat, value)
- Rule:
    - Remove all invalid values (-999).
    - value is the observed temperature.

2.3 Implementation

The dataset was parsed from the XML file and reshaped into a geographic grid. Longitude and latitude were calculated as:

$$lon = 120.00 + 0.03 \times col, \quad lat = 21.88 + 0.03 \times row$$

Two DataFrames were created:

- **df_clf** = classification dataset (lon, lat, label)
- **df_reg** = regression dataset (lon, lat, value)

3. Models and Training

3.1 Classification Model

- Model: **Random Forest Classifier** (200 trees)
- Problem: Severe imbalance between valid and invalid points.
- Solution: **Random downsampling** was applied to balance the dataset.

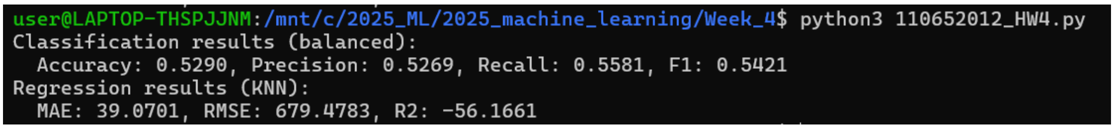
3.2 Regression Model

- Model: **K-Nearest Neighbors (KNN, k=5)**
- Motivation: KNN is suitable for spatial interpolation, predicting a grid point’s value by averaging its nearest neighbors.

3.3 Evaluation Metrics

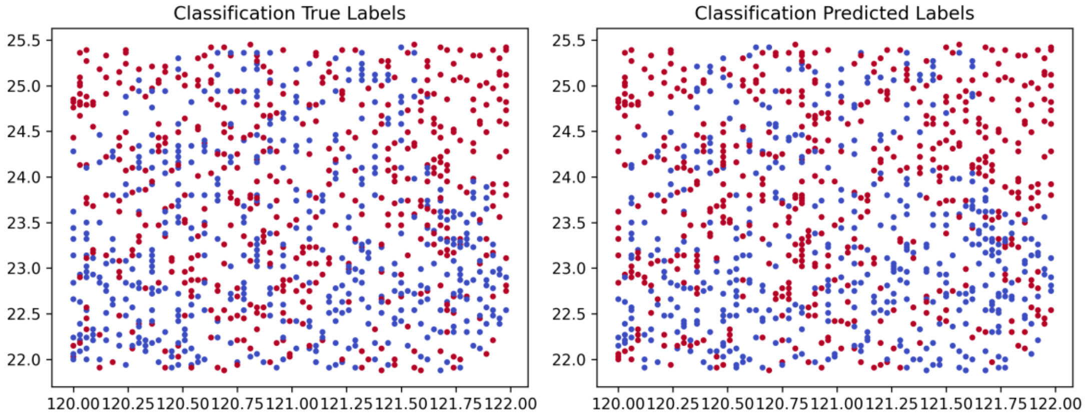
- **Classification**: Accuracy, Precision, Recall, F1-score.
- **Regression**: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2.

4. Results

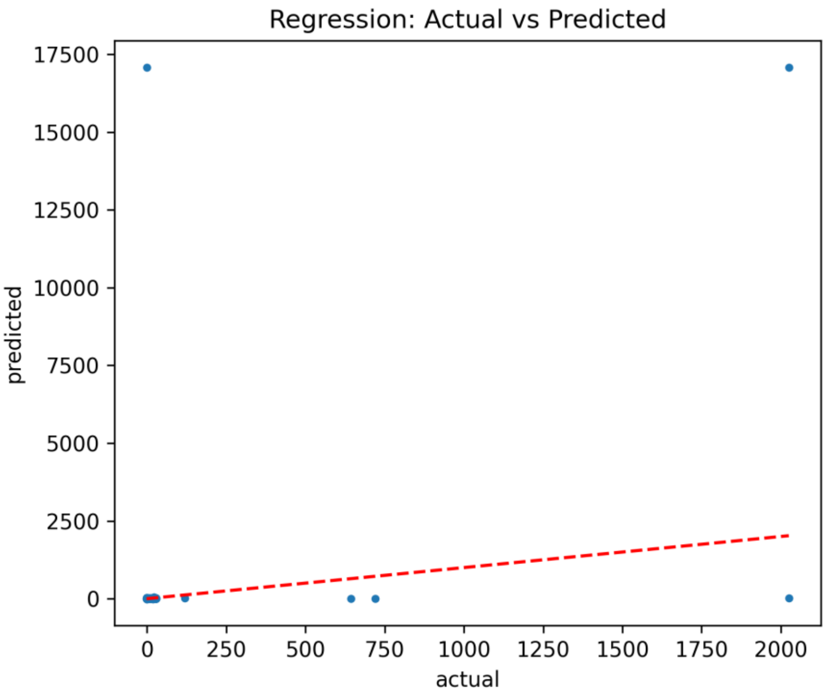


```
user@LAPTOP-THSPJJNM:/mnt/c/2025_ML/2025_machine_learning/Week_4$ python3 110652012_HW4.py
Classification results (balanced):
  Accuracy: 0.5290, Precision: 0.5269, Recall: 0.5581, F1: 0.5421
Regression results (KNN):
  MAE: 39.0701, RMSE: 679.4783, R2: -56.1661
```

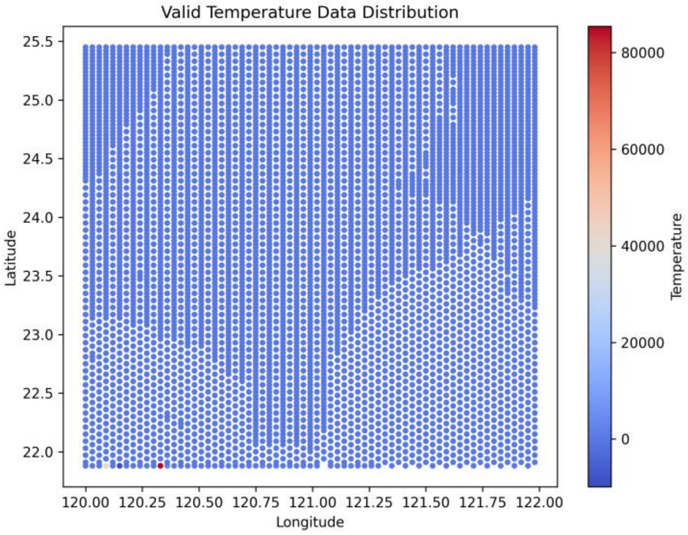
4.1 Classification results:



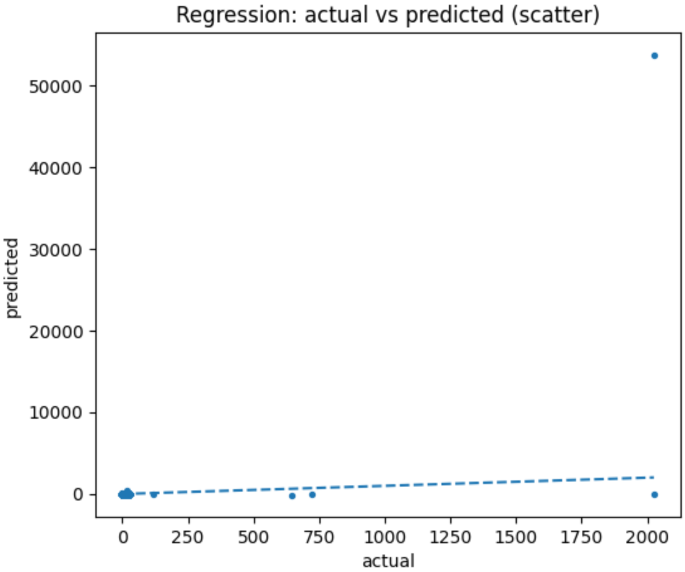
4.2 Regression scatter:



4.3 Heatmap of valid temperature data:



4.4 result:



5. Discussion

The results show that both tasks were strongly limited by the dataset. For classification, the extreme imbalance between valid and invalid points meant that the Random Forest model could only reach about 53% accuracy even after balancing. This indicates that longitude and latitude alone do not provide enough information to separate valid and invalid points.

For regression, the KNN model performed poorly with high error values and a strongly negative R2R^2R2. This reflects the fact that temperature cannot be explained by coordinates alone. The sparse distribution of valid points further reduced the model’s ability to generalize, making interpolation unreliable.

6. Conclusion

This assignment successfully demonstrated the transformation of XML data into classification and regression datasets, and the application of machine learning models to both tasks. However, the models achieved only limited performance due to data sparsity and insufficient features. Future work should focus on adding richer meteorological variables or applying interpolation methods to improve prediction accuracy.