AI CUP 2024 秋季賽

根據區域微氣候資料預測發電量競賽報告

**隊伍：TEAM\_5709**

隊員:林靖鈞（隊長）

Private leaderboard:310872.06/Rank 1

**壹、環境**

1. 開發工具與框架

本次實驗使用 Python 3 語言，並採用了以下主要套件進行數據處理與模型構建：

套件：numpy==1.24.3

pandas==1.5.3

matplotlib==3.7.1

seaborn==0.12.2

yellowbrick==1.5

scikit-learn==1.2.2

xgboost==1.7.6

lightgbm==3.3.5

catboost==1.2

開發環境為jupyter notebook並構建於 Ubuntu 24.04 上，使用 Conda 管理虛擬環境，並記錄所需的套件及其版本以確保實驗的可重現性。

2. 硬體環境

本次競賽中，本隊使用了兩台機器來完成整個實驗，分別用於特徵工程與模型訓練，具體配置如下：

操作系統：Ubuntu 24.04 Server（伺服器） / Ubuntu 24.04（筆電）

處理器：AMD Ryzen 7 8845HS（伺服器） / AMD Ryzen 5 4600H（筆電）

記憶體：64GB（伺服器） / 32GB（筆電）

顯示卡：無（兩台機器均未配置 GPU）

硬體資源的分工設計如下：

筆電主要用於特徵工程與數據分析，因其便攜性便於隨時檢視與調整數據處理流程。

伺服器則用於訓練模型，依賴其更高的計算性能和大容量記憶體以支持大規模數據運算。

3. 額外數據集

在本次競賽中，本隊的重點主要放在對官方提供的數據集進行深入的特徵工程。除了異常值檢測與數據處理之外，本隊的實際訓練資料完全基於官方數據集進行建模，並未引入任何外部數據。訓練資料的特徵構建主要依賴於太陽位置的物理模擬，通過對數據的物理特性和規律進行深度挖掘來提升模型的預測能力。

**貳、演算方法與模型架構**

這次比賽本隊設計了一個基於堆疊回歸（Stacking Regressor）的模型架構，如下圖1所示

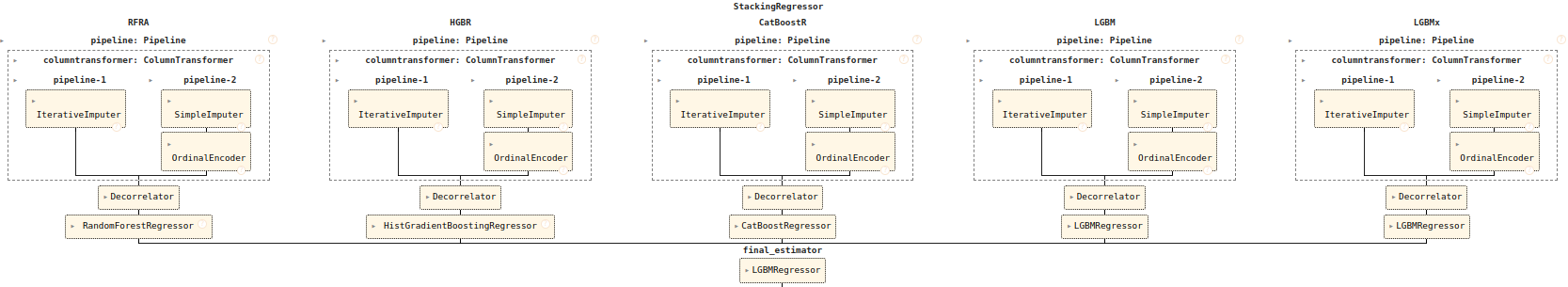


圖1: 第一層包含了多種基礎回歸模型; 第二層選擇了LightGBM回歸器作為最終預測器

1. 第一層（Level 0）

第一層**包含多個基礎回歸模型**，每個模型針對不同的數據特性進行學習，從而捕捉數據的多樣性與複雜性。這些模型包括：

* RandomForest (RFRA)：使用絕對誤差（absolute\_error）作為損失函數，並通過 1000 棵決策樹學習數據特徵。其優勢在於能有效處理非線性模式，並對異常值具有較高的魯棒性。
* HistGradientBoosting (HGBR)：通過直方圖方法進行梯度提升，顯著提升了訓練效率，並同樣採用了絕對誤差作為損失函數。為避免過度擬合並加速訓練，啟用了早停機制（early\_stopping=True）。
* CatBoost Regressor (CatBoostR)：這是一種高效的梯度提升模型，內建對類別型數據的支持，且對數據排列順序不敏感，尤其適合異構數據建模。
* LightGBM (LGBM) 和 LGBMx：兩個配置不同的 LightGBM 模型。LGBM 以均方誤差（regression）為目標函數，適用於接近正態分佈的數據；而 LGBMx 以平均絕對誤差（mae）為目標函數，能更有效處理異常值。兩個模型均設置了 15,000 次迭代（n\_estimators=15000）和 0.15 的學習率（learning\_rate=0.15），並啟用了多線程處理（n\_jobs=-1）。

2. 第二層（Level 1）

在第二層本隊選擇了一個 LightGBM 回歸器作為最終的預測器。該回歸器同樣以平均絕對誤差（MAE）為目標函數，並引入了額外的隨機性（extra\_trees=True），進一步提升模型的泛化能力。最終模型配置包括 10,000 次迭代（n\_estimators=10000）和 0.1 的學習率（learning\_rate=0.1），以確保模型的穩定性與精確性。

**參、創新性**

在本次比賽中，本隊本隊在資料處理與模型應用上進行了一系列創新設計，特別是在數據集擴建與預測策略方面，突破了傳統時間序列模型的限制。

1. 交換天氣特徵的方法

本隊提出了一種獨特的數據擴建策略—交換天氣特徵：通過將兩個站點的數據進行交換，目標變數（Power）保留原站點的資料，而天氣特徵則使用另一站點的數據。這種方法突破了單站點數據的局限性，充分挖掘了多站點天氣數據的相關性，顯著擴充了訓練集的數量（從 11 萬筆提升至 75 萬筆）。此策略不僅提升了數據的多樣性，還使模型能學習到更加穩定且普適的特徵，有效提高了模型對不同天氣條件的適應能力。

2. 靈活的時間粒度應用

本隊創新地結合了不同時間粒度的數據，根據模型訓練與預測的需求進行靈活調整。在訓練階段，採用 每十分鐘 的數據進行模型學習，這種降頻處理大幅降低了數據噪音，提升了訓練穩定性與效率；而在預測階段，使用 每分鐘 的高精度數據，並對其輸出進行平均處理。這種策略在保證數據精細度的同時，捕捉到了細微的時間變化特徵，進一步提升了預測的準確性。

3. 測試集數據擴充

針對測試集，本隊將測試數據恢復為 每分鐘 粒度，並將多個站點的天氣特徵整合為輸出特徵，最終通過對輸出序號進行平均處理，將測試集數據量增加至 720,892 筆。這一處理方法充分利用了多站點數據的補充效應，使模型對測試集的感知更加全面，從而顯著提升了預測的穩定性和精度。

**肆、資料處理**

**資料處理策略**

訓練集和測試集的資料處理分為三個主要階段：異常值處理 數據集擴建 和 測試集處理。本隊根據比賽需求對原始資料進行了多次優化，確保數據的有效性與模型的適用性。

**1. 異常值處理**

本隊根據氣象局發布的颱風與豪大雨特報，對數據中可能存在異常的部分進行剔除，共去除了 3235 筆原始資料。然而，在後續分析中發現，測試集中包含兩天與颱風相關的天氣情境。為了更準確地模擬這些情境，本隊調整了處理策略：針對颱風的兩天數據，選取原始資料中的正常數據 10%，並結合先前剔除的 3235 筆資料，對氣壓數據進行標準化處理，進一步基於這些數據進行額外推理，從而強調颱風天氣對太陽能發電的影響。

**2. 數據集擴建與特徵工程**

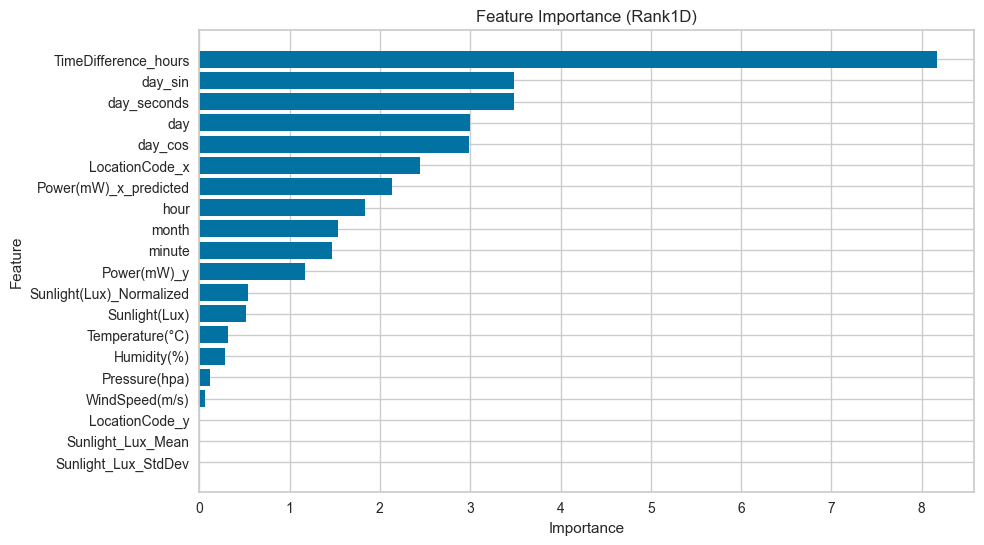
在進行模型訓練時，本隊發現使用分鐘級數據會引入過多噪音，從而影響模型的穩定性。因此，本隊將原始的約 110 萬筆分鐘級資料降頻至 10 分鐘精度。這樣的處理不僅與欲輸出測試集的格式保持一致，還有效降低了數據的噪音水平，經降頻處理後，數據量縮減為約 11 萬筆。

為進一步擴充訓練數據，本隊採用了 「交換天氣特徵」 的方法：選擇兩個不同的站點，將它們同一時間的數據進行交換——目標變數（power）依然使用原站點的資料，而天氣特徵則使用另一站點在同一天的數據。這種方法將數據量從原本的 11 萬筆大幅提升至約 75 萬筆，有效增強了數據的多樣性，並進一步提升了模型在不同天氣條件下的泛化能力。

此外，本隊還引入了多項新增特徵，進一步提升了數據的表達能力(圖2)：

* **Power(mW)\_x\_predict：來自堆疊模型上一次預測過程的輸出結果，作為特徵參數之一，用於改進最終模型的表現。**
* 時間相關特徵：包括時間差特徵（TimeDifference\_hours）以及週期性特徵（day\_cos 和 day\_sin）。這些特徵捕捉了數據與太陽日週期的相對位置以及一天內的週期性變化。
* 時間分量特徵：從 DateTime 欄位中提取的分鐘、時、日、月特徵（Minute、Hour、Day、Month），幫助模型識別日內與月內的特定規律。
* 光照數據特徵：針對 Sunlight(Lux) 數據計算的標準差（StdDev）、平均值（mean）和正規化值（normalized），雖然提升有限，但有助於捕捉光照條件的變化。

圖2: 訓練用的特徵重要性排列



**3. 測試集處理**

針對測試集，本隊採用了不同的擴充策略：將測試資料恢復為分鐘級數據，並將多個站點的天氣特徵作為輸出特徵，最終按照序號對預測結果進行平均。這一處理將測試集的數據量提升至 720,892 筆，大幅增加了模型對測試集的感知能力與穩定性，從而顯著提升了預測精度。

**伍、訓練方式**

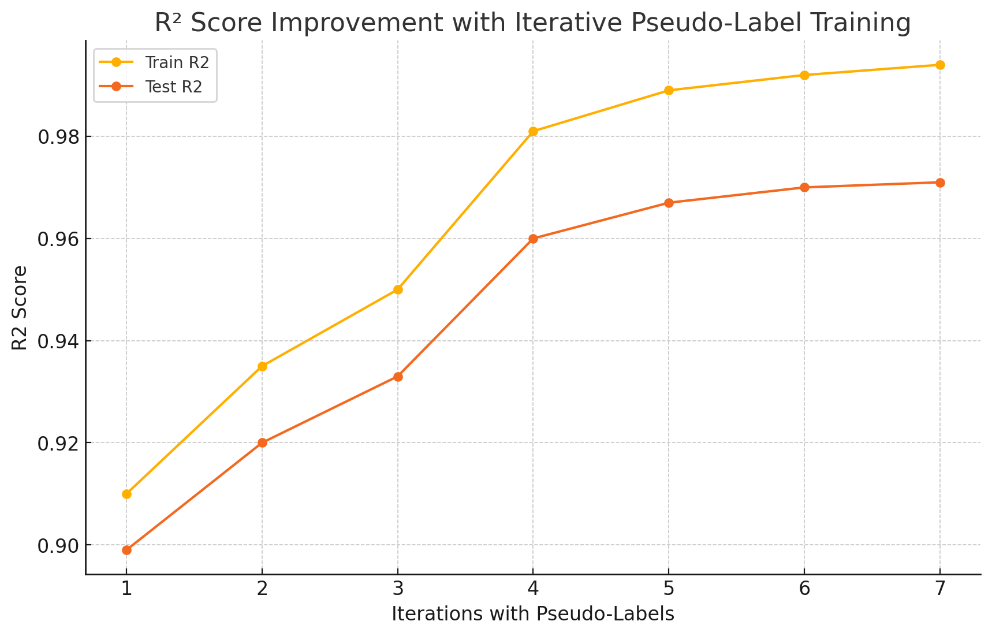
本隊採用了 **偽標籤（Pseudo-Label）**迭代技術，用於進一步提升模型的性能。該技術通過將上一輪訓練的預測結果作為下一輪訓練的偽標籤，不斷優化模型的準確性。在實驗中，本隊觀察到以下趨勢：

1. 在第 3 至 4 次迭代 時，模型的 R² 分數提升最為顯著，特別是對測試集（Test R²）的準確性提升效果尤為明顯。
2. 在第 5 次迭代 後，R² 分數趨於平緩，但仍有輕微提升。
3. 在第 7 次迭代 時，本隊決定停止迭代。原因如下：

* 訓練集與測試集的 R² 分數差距逐漸增大，可能出現過擬合現象。
* 訓練時間顯著增加，但性能提升有限。

如圖3所示，透過這一技術，本隊在平衡訓練效率與模型性能之間達到了較佳的效果，使得模型在測試集上的表現更加穩定且精確。

圖3: 偽標籤（Pseudo-Label）多次迭代改善效果



在本次比賽中，官方的Public LB提到使用了部分數據進行評估，但具體使用了多少筆資料並不清楚，這使得本隊在本地驗證中遇到了以下問題：

1. 本地驗證的 MAE（平均絕對誤差） 與提交結果的總誤差（Public LB）表現不一致。
2. 出現了本地 MAE 降低但提交後總誤差未改善的情況。

為了解決這一問題，本隊改變了評估策略，選擇 R² 分數 作為模型的主要評估指標，原因如下：

* R² 分數的上升通常伴隨著總誤差的下降，這與比賽的最終目標一致。
* R² 能更好地捕捉模型對整體趨勢的擬合效果，而不僅僅是針對絕對誤差的最小化。

**最終，通過使用 R² 分數作為核心評估指標，本隊能更有效地調整模型參數並提升預測結果的準確性，從而在比賽中取得更穩定的表現。**

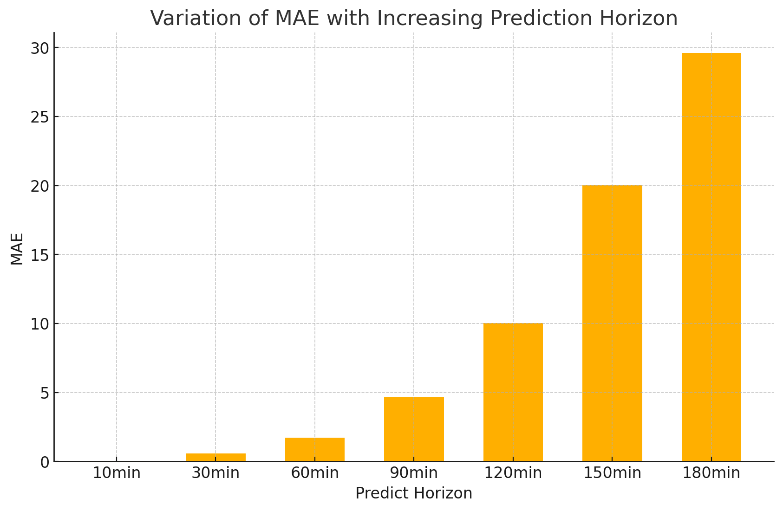
**陸、分析與結論**

**1.分析**

實際拿到題目之前，我參考了許多針對中短期時間序列預測（如風力、太陽能等）的研究方法，其中使用 CNN-LSTM 結合的模型相當普遍。因此，本隊的本隊一開始也採用了這種方法，並在測試中獲得了 public 成績 512,323。然而，這樣的表現顯然不足以讓本隊在本次競賽中脫穎而出。

通過對結果進一步分析後，本隊發現模型的預測誤差會隨著預測時間範圍的延長呈指數級上升(圖4)，導致在更長時間範圍內的預測效能表現不佳。這一現象顯示了本隊現有方法在解決時間序列長期預測問題時的局限性。這意味著要本隊要找尋新的方法來提昇預測精度。

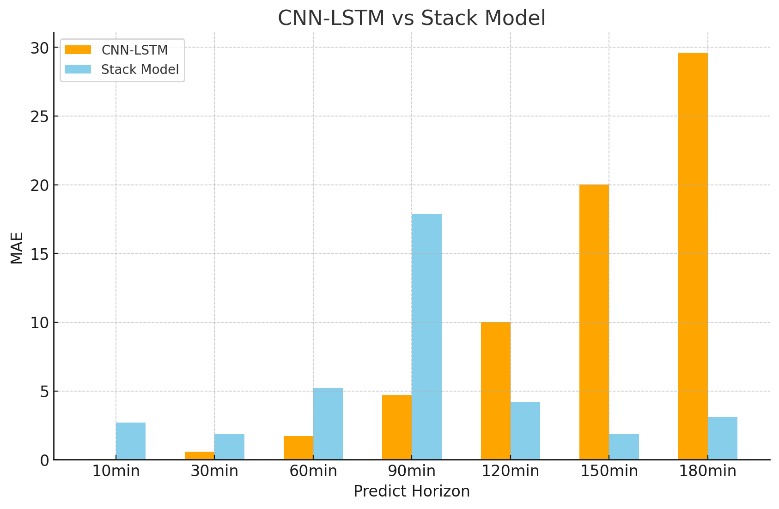
圖4: CNN-LSTM會隨時間範圍的延長呈指數級上升



在重新建模的過程中，結合以往參加比賽與進行 side project 的經驗，我認為 tabular regression 方法能有效避免因預測時間跨度過長而導致的精度下降問題。因此，在本次比賽中，我特別選擇了多種模型組合的堆疊模型（stack model），包括 HistGradientBoosting、LightGBM、XGBoost、RandomForest 和 CatBoost，並在堆疊模型的第二層選擇 LightGBM 作為最終的集成模型。

這些模型分別學習了在不同氣象條件下，天氣特徵與發電量之間的關係，充分發揮了各算法的特點與優勢，從而顯著提升了整體預測的準確性。儘管在預測前 90 分鐘與極端值時，堆疊模型的誤差略高於 CNN-LSTM，但整體而言，堆疊模型在絕對總誤差上的表現更加優異。

圖4: CNN-LSTM與Stack Model MAE的比較



在此基礎上，本隊進一步嘗試引入 偽標籤（pseudo-labeling） 技術。該技術的核心在於利用未標註的數據進行預測，並將預測結果作為新特徵加入訓練數據集中，以進一步強化模型的性能。實驗結果顯示，加入偽標籤後，模型的準確性確實得到了進一部的提升。

1. **結論**

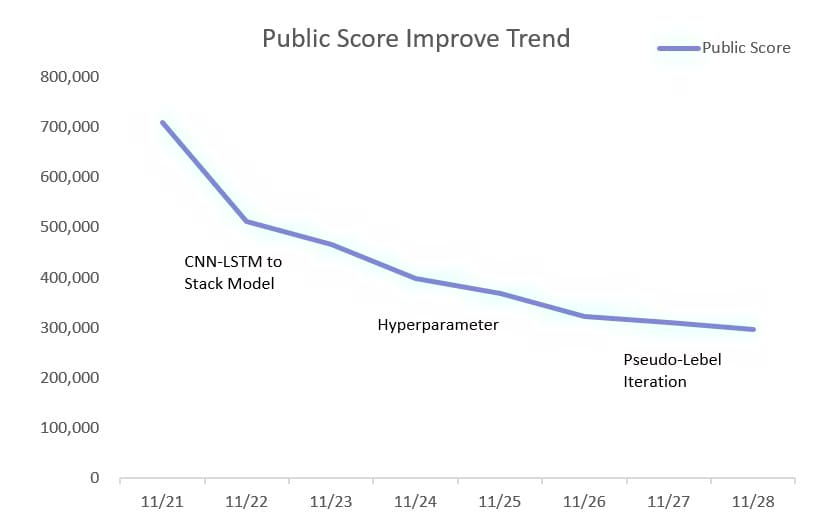
本次競賽中，我們僅使用官方提供的資料，即獲得 public 與 private 排行榜第一。本隊研究中發現，若加入更精準的落塵量、地表溫度等環境變數，便可大幅提升預測模型的表現，這顯示，未來在數據收集中適時補充與環境相關的關鍵資訊，對於提升模型的準確度與穩健性至關重要。

針對新設站點的天氣交互特徵設計策略，我們設計了新設站點的天氣交互特徵，模擬同日異地的氣象條件，讓模型學習跨地點氣候變化對發電量的影響。此舉有效拓寬了模型學習空間，增強泛化能力與適應性。實驗顯示此策略不僅提升本次太陽能發電量預測精度，亦在更複雜場景中展現優良表現。

更關鍵的是，此方法可應用於無歷史數據的新電廠，僅須他區氣象與發電數據，即能準確預估，為新電廠選址與效能評估提供科學依據。同時，我們認為此方法對風力發電預測、城市氣象模擬、跨區域農業生產分析等領域的時空序列建模同樣有啟示。

未來，隨著數據來源的多樣化與精細化，若能進一步結合遙測數據、衛星影像與更豐富的地面監測信息，該模型有望在預測跨區域能源管理與可再生能源分配中發揮更大的作用，為台灣的再生能源發展提供穩定且強大的預測結果。

圖5:最後七天Public LB進步的主要方法



**柒、程式碼**

Github連結：https://github.com/xup6193193/AICUP\_2024\_solar\_1128/tree/main

**捌、使用的外部資源與參考文獻**

* Tovar, M., Robles, M., & Rashid, F. (2020). PV power prediction, using CNN-LSTM hybrid neural network model: Case of study: Temixco-Morelos, México. *Energies, 13*(6512). https://doi.org/10.3390/en13246512
* Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems (NIPS)*.
* Zhang, Y., & Wang, L. (2022). A comprehensive review of battery management systems for electric vehicles. *Journal of Power Sources, 523*, 231–247. <https://doi.org/10.1016/j.jpowsour.2022.231247>
* Zhang, Y., & Wang, L. (2023). A comprehensive review of battery management systems for electric vehicles. *Energy Storage Materials, 45*, 123–145.
* Billet, P. (2022). Solar power generation, stacking and importances [Kaggle notebook]. *Kaggle*. <https://www.kaggle.com/code/philippebillet/solar-power-generation-stacking-and-importances>

**額外資料集**

中央氣象局資料： [https://ci.taiwan.gov.tw/dsp/Views/dataset/weather.aspx](https://disq.us/url?url=https%3A%2F%2Fci.taiwan.gov.tw%2Fdsp%2FViews%2Fdataset%2Fweather.aspx%3AgZdAdntMh7SL2l9tBsSvScyp2P0&cuid=7636807)

環境署空氣品質資料： [https://data.moenv.gov.tw/dataset/detail/AQX\_P\_35](https://disq.us/url?url=https%3A%2F%2Fdata.moenv.gov.tw%2Fdataset%2Fdetail%2FAQX_P_35%3A0Rs9XPw5UcMZmG7rmrXxoNd9rmY&cuid=7636807)

Himawari 衛星遙測資料： [https://registry.opendata.aws/noaa-himawari](https://disq.us/url?url=https%3A%2F%2Fregistry.opendata.aws%2Fnoaa-himawari%2F%3ARYbZw6MtoApFZiI5wk_Kn-pKisE&cuid=7636807)

農業氣象觀測資料：https://agr.cwa.gov.tw/NAGR/

歷史天氣警報：https://opendata.cwa.gov.tw/promotion/introduction/warning

作者聯絡資料表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 表X: (此處加上描述)  隊伍名稱 | TEAM\_5709 | Private Leaderboard 成績 | 310872.0 | Private Leaderboard 名次 | 1 |
| 身分  （隊長/隊員） | 姓名 (中英皆需填寫)  (英文寫法為名,姓，例：Xiao－Ming, Wu，名須加連字號，姓前須加逗號) | 學校＋系所中文全稱  (請填寫完整全名，勿縮寫) | 學校＋系所英文中文全稱  (請填寫完整全名，勿縮寫) | 電話 | E-mail |
| 隊長 | 林靖鈞  Lin-jing-jun | 國立竹北高中普通科 | National Chupei Senior High School | 0906773196 | xup6vu84m3ul6@gmail.com |
| 隊員1 |  |  |  |  |  |
| 隊員2 |  |  |  |  |  |
| 隊員3 |  |  |  |  |  |
| 隊員4 |  |  |  |  |  |

★註1：請確認上述資料與AI CUP報名系統中填寫之內容相同。自2023年起，獎狀製作將依據報名系統中填寫內容為準，有特殊狀況需修正者，請主動於報告繳交期限內來信moe.ai.ncu@gmail.com。報告繳交截止時間後將不予修改。

★註2：繳交程式碼檔案與報告，請Email至：ailabailab5051@gmail.com，並同時副本至：t\_brain@trendmicro.com與moe.ai.ncu@gmail.com。缺一不可。