基於機器學習實現智慧化水質分析與評估系統

(一) 研究計畫之背景。

1. 研究計畫之背景

本研究旨在開發並實現具機器學習（Machine Learning, ML）功能的智慧化水質分析與評估系統，以解決目前水質資料分析缺乏高效率方法的問題。研究作者曾擔任科技部計畫兼任研究助理，協助執行「運用智慧型綠能永續景觀噴泉促進環境保育效益之研究」（計畫編號：MOST 110-2410-H025-024），該計畫運用物聯網（Internet of Things, IoT）技術蒐集南投麒麟潭的水質資料。藉由該計畫經驗發現自動化運行的結果相較於人工記錄更有效率，但後續資料分析仍需耗費大量工作時間。

2. 研究問題

在水質資料分析方面，除了使用 Microsoft Excel 直接計算外，通常會使用如 Statistical Analysis System（SAS）或 Statistical Product and Service Solutions（SPSS）等統計專有軟體（Proprietary Software）。然而，這些軟體為專有（Proprietary）且非開源，不僅需要額外購買授權，還需投入時間學習才能熟練使用。由於其操作高度依賴使用者的專業知識，分析結果可能因個人經驗而有所差異。由於這些軟體無法修改與物聯網設備直接進行資料串聯傳輸，造成未來計畫整合實現自動化資料收集與資料分析的障礙。

3. 研究原創性

近十年來，人工智慧（Artificial Intelligence, AI）從理論發展到實際應用，各種 AI 等工具也被廣泛運用於日常問題成為解決方案 [1]。本研究運用人工智慧中的機器學習技術，開發一套智慧化水質分析與評估系統，實現更高效且易於操作的創新水質分析方法。不同於傳統統計軟體需要人工輸入與分析，本系統將結合自動化資料處理與機器學習演算法，以提升水質分析的準確性與效率。此外，本研究將探索 AI 在水資源管理與環境保育領域的創新應用，提供創新的智慧化水質資料分析方案。

為了找出最適合水質分析的機器學習模型，本研究針對 XGBoost、LightGBM、支援向量機（Support Vector Machine, SVM）、隨機森林（Random Forest, RF）、多元多項式迴歸（Multiple Polynomial Regression, MPR）以及線性迴歸（Linear Regression, LR）等多種模型進行比較與評估。透過實驗分析與比較不同模型在水質資料分析準確度、計算效能與資料適應性的表現，進一步應用於改良水質分析，使系統能夠有效應對水體環境的分析與評估需求。

實驗設計包括使用先前於「運用智慧型綠能永續景觀噴泉促進環境保育效益之研究」計畫中，使用物聯網設備實際蒐集的水質資料，加上環境部環境資料開放平臺公開資料使用，特徵選擇包括溶解氧（DO）、生化需氧量（BOD）、氨氮（NH3-N）、電導度（EC）和懸浮固體（SS）等關鍵指標。在訓練過程中，使用水體品質指標（Water Quality Index, WQI_5）作為標記資料進行監督式學習。

4. 研究重要性

在過去十年內，人工智慧的研究從長期被忽視的陰霾中走了出來，被人們看見了希望且掀起了新一波風潮 [2]，而水資源管理與環境生態保育是全球關注的議題，而水質分析與評估是確保水資源安全的重要環節。本研究透過機器學習技術，克服傳統方法的效率瓶頸，降低對專業知識依賴，使水質資料分析變得更有效且易於操作，提升水質資料分析技術的普及性與可行性。

本研究針對 XGBoost、LightGBM、支援向量機（SVM）、隨機森林（RF）、多元多項式迴歸（MPR）以及線性迴歸（LR）等機器學習模型進行獨立訓練，並透過決定係數（R^2）、均方根誤差（Root Mean Square Error, RMSE）、平均絕對誤差（Mean Absolute Error, MAE）以及殘差分析等指標評估各模型在水質資料分析上的準確性。

透過實驗分析，我們發現〔TODO〕。這些結果有助於選擇最適合水質資料分析的機器學習模型，進一步提升水資源管理的準確度，為環境監測技術與未來相關研究的發展提供有效的參考依據。

5. 預期影響性

本研究的智慧化水質分析與評估系統結合了模型比較實驗與系統開發，有望在以下方面產生影響：

1-5-1. 技術創新：人工智慧未來將成為人類不可或缺的夥伴 [3]，本研究在水質資料分析引入機器學習技術，提升分析準確性與自動化程度。

1-5-2. 環境保育：能夠在水資源管理方面準確且有效地發現水質變化，改善水體生態環境。

1-5-3. 跨領域應用：本研究結合了資訊工程、人工智慧、水資源管理和環境生態保育等不同領域，未來也可擴展應用於不同環境監測領域，例如空氣品質監測或污染物分析，促進人工智慧在環境科學中的應用。

6. 國內外研究情況

本研究作者先前參與的計畫已經使用物聯網技術進行水質資料的採集，這種方法在環境監測中已經相當普遍。然而，目前國內外針對水質資料分析的智慧化系統，特別是結合機器學習技術的水質分析系統相對較少。機器學習技術在水質評估中的應用已被證明是有效的，並逐漸成為水資源管理的重要分析工具 [4]，應用各種技術構建學習模型，能夠達成提供即時準確水質資訊的目的，期望得以透過早期診斷水質異常以改善公共健康 [5] 因此，本研究具有創新性和獨特性，有助於填補水質資料分析中高效方法的空白。

7. 重要參考文獻

隨著人工智慧發展，各種機器學習模型也如雨後春筍般出現。模型實作上如開發一種基於多元多項式迴歸（MPR）的機器學習模型，在該研究中以模型與線性迴歸及支援向量機模型進行比較，發現 MPR 在性能評估指標包含 R^2（R-squared）和 RMSE（Root-Mean-Square Error）上皆優於另外兩個模型 [6]。加上其他研究成果使用 MPR 模型驗證其預測結果 [7]，MPR 建立了一個六元多項式迴歸模型，結果顯示其訓練集的 MAPE（Mean Absolute Percentage Error）為 1.873%，MAE（Mean Absolute Error）為（TODO 待確認），證明了 MPR 模型在水質資料分析上的優越性 [8]。這些研究成果將有助於本研究的模型選擇與效能評估，也因此本研究前期採用 MPR 模型作為基準模型進行比較。

另外，隨著近年來 XGBoost 與 LightGBM 等新興模型在機器學習領域中的崛起，本研究在中後期將其納入比較範疇，XGBoost 在資料科學領域中被廣泛使用，也在 Kaggle 許多機器學習及競賽中取得最佳結果，其特點為有非常強的擴展性（Scalibility）與性能，可以使用相對其他模型更少的系統資源，擴充數十億級別的資料 [9]。而 LightGBM 是一種高效能的梯度提升決策樹，在傳統 Gradient Boosting Decision Tree（GBDT）演算法上加入使用 Gradient-based One-Side Sampling（GOSS）和 Exclusive Feature Bunding（EFB），並顯示出 LightGBM 相較傳統的 GBDT 加快了 20 倍以上的速度，同時也並未降低準確性 [10]。

綜合以上重要參考文獻，本研究對 XGBoost、LightGBM、支援向量機（Support Vector Machine, SVM）、隨機森林（Random Forest, RF）、多元多項式迴歸（Multiple Polynomial Regression, MPR）以及線性迴歸（Linear Regression, LR）等多種模型進行比較與評估，期望找出最適合水質分析的機器學習模型，並提升水質分析的準確性與效率。

(二) 研究方法、進行步驟及執行進度。

1. 研究方法

本研究開發之智慧化水質分析與評估系統採用前後端分離架構，透過應用程式介面（Application Programming Interface, API）進行資料傳遞與呼叫。後端（Back-end）水質資料分析模型以 Python 程式語言開發，運用 NumPy、Pandas 及 Scikit-learn 等工具進行水質資料的處理和分析。前端（Front-end）則以 JavaScript 語言的 React Native 框架開發行動裝置應用程式，實現跨系統平臺操作。此架構的水質資料分析模型開發可歸納為四個步驟：匯入資料、資料處理、特徵工程及模型訓練與評估。系統建置完成後，使用者在前端行動裝置應用程式中只需輸入欲分析的水質資料，即可經由已訓練完成的機器學習模型進行分析，再於使用者介面（UI）將產生一份分析報表，包括水質狀況的綜合評分以及狀態說明。

在模型開發過程中，我們認為機器學習技術不僅適用於水質資料分析，也能拓展到其他相關領域。除了前期使用的 MPR 模型外，我們還加入 XGBoost 和 LightGBM 等新興優秀模型進行比較，以尋找最適合解決問題的方案。同時，利用交叉驗證（Cross Validation）調整超參數（Hyperparameter），有效防範過擬合並評估模型的泛化能力 [11]。研究工具方面涵蓋了機器學習程式語言 (2-1-1)、開發環境 (2-1-2)、功能函式庫 (2-1-3) 及前端開發工具 (2-1-4)，用以構建一套完整的水質資料分析與評估系統，並確保能挑選出最佳模型。各項工具分述如下：

2-1-1. 機器學習語言

近年來 Python 在人工智慧領域中具有顯著影響力，因其優秀的資料處理與分析能力，支援各式開源函式庫使得 Python 成為資料分析的熱門選擇 [12]，且能夠有效處理即時、大型和非結構化資料 [13]，對於本研究進行各式水質資料分析的研究具有明顯的優勢。

2-1-2. 機器學習環境

本研究機器學習環境主要分為 (A) 雲端運算平臺及 (B) 開發環境。正如所言「工欲善其事，必先利其器」，為了能夠有效率地進行整體系統的各項開發，並在完成後部署於雲端運算平臺。

A. 雲端運算平臺

本研究運用機器學習框架－ Scikit-learn 進行功能開發，Scikit-learn 著重高效能的運算資源利用，符合智慧化水質資料分析的應用需求。此外，亦將建置雲端平臺－開發與部署所建置之 MPR 模型於 Google Cloud Platform（GCP），GCP 為一種基礎結構即服務（IaaS）平臺，提供於雲端部署機器模型之能力。

B. 開發環境

本研究採用 Visual Studio Code 作為整合開發環境（Integrated Development Environment, IDE），並透過 Docker 實現容器化開發。可以直接將所開發之機器學習模型與應用程式部署至雲端平臺，運用分散式架構使得本研究具備處理和運算大量水質資料的能力。

2-1-3. 功能函式庫

程式開發所需之功能函式庫分為：(A) 資料處理工具、(B) 機器學習框架、(C) 圖表繪製工具及 (D) 應用程式介面。實作上以模組化方式結合各項工具能有助於完成水質分析模型的建構，相關說明如下：

A. 資料處理工具

本研究以 NumPy 及 Pandas 作為資料處理工具，NumPy 是 Python 程式語言中支援科學和數值計算的函式庫，也相容於 Pandas、Matplotlib 和 Scikit-learn 等工具 [14]。NumPy 作為資料處理的主要工具是因為它提供了科學計算所需的基本資料結構 [15]。同時 NumPy 也提供豐富的數學運算、資料讀寫、線性代數、傅立葉變換和隨機數生成等功能函式 [16]。使用 NumPy 得以更有效率地處理原始資料，進一步提升程式的開發效率。而 Pandas 建立於 NumPy 的基礎之上，是一種用於格式化資料分析的開源 Python 函式庫，它能夠處理類似電子試算表格式的輸入資料，可以快速地進行資料載入、操作、對齊和合併等功能 [17]。

B. 機器學習框架

本研究採用了 Scikit-learn 作為機器學習框架，以建構研究所需的各種機器學習模型模型 [18]。支援 Python 程式的機器學習框架之中，Scikit-learn 提供多樣的演算法，其特色包括具有良好的應用程式介面（Application Programming Interface, API）、詳細文檔說明及技術支援 [19]。由於 Scikit-learn 具有輕量而高效的特性，可使本研究實作能夠有效率地達到智慧化水質資料分析的應用需求。

C. 圖表繪製工具

Matplotlib 是一個多功能的 Python 繪圖函式庫，具有產生各種類型和格式圖表的功能，包括折線圖、散布圖、熱圖、直方圖、圓餅圖和立體圖形，同時也支援動畫和互動顯示 [20]。運用 Matplotlib 能夠以視覺化方式呈現多元多項式迴歸運算後的水質資料分析結果，使模型訓練結果呈現更生動以提升可讀性。

D. 應用程式介面

本研究以 FastAPI 作為應用程式界面（API）框架，FastAPI 是一個新興但可靠的 API 框架 [21]，能夠快速且有效地於 Python 建立 RESTful API 伺服器。機器學習模型與使用 FastAPI 服務的前端整合，能夠相較使用如 Flask 等其他 API 框架的效能提高了 45% [22]。使用 FastAPI 建立 API，有助於前後端之間的通訊與資料傳遞，實現智慧化水質分析與評估系統的系統化架構。

2-1-4. 前端開發工具

在進行前端應用程式開發的過程中，本研究採用 React Native 框架。此框架由 Meta 公司基於 React 框架與 JavaScript 語言設計開發，重點在於能夠開發跨操作系統

運行的行動裝置應用程式，作業系統支援包括 Android 及 iOS [23]。使用 React Native 不僅能夠直接進行跨系統平臺開發，還能顯著提升行動裝置應用程式開發的速度。使得本研究在開發具機器學習功能之智慧化水質分析與評估系統時，前端開發的效率明顯優於單獨針對 Android 或 iOS 進行開發的方法。

2. 研究步驟

本研究所設計之多元多項式迴歸模型開發與實驗步驟流程如圖所示，可分為 (2-2-1) 資料取得、(2-2-2) 資料前處理、(2-2-3) 模型選擇與比較、(2-2-4) 超參數調整、(2-2-5) 改良與測試及 (2-2-6) 實現系統化，整體系統化作業流程包含起始於資料取得、模型完成與最終實際應用於水質分析，以實現「基於機器學習實現智慧化水質分析與評估系統」為目標。

2-2-1. 資料取得

本研究將參考「運用智慧型綠能永續景觀噴泉促進環境保育效益之研究」計畫（編號：MOST110-2410-H025-024）所蒐集的水質資料，並結合中華民國環境部全國環境水質監測資訊網的公開資料，共計蒐集了 87,005 筆水質資料。本研究所採用的水質參數包括溶氧量（DO）、生物需氧量（BOD）、懸浮固體（SS）、氨氮（NH3-N）及導電度（EC）。研究中將使用多元多項式迴歸（MPR）模型對各項水質資料進行機器學習分析。如圖所示，不同資料之間存在交互影響水質健康的關係，例如：DO 濃度越高，NH3-N 及 EC 濃度越低；NH3-N 濃度越高，EC 越高。此外，DO 與水質綜合評分之關係顯示其為重要指標之一，而 BOD、NH3-N、EC 及 SS 與水質綜合評分之關係則顯示其為水質污染的指標。

2-2-2. 資料前處理

在進行機器學習模型訓練前，需對蒐集的原始資料進行前處理，以減少母體資料中的不確定性。此過程可以透過 Pandas 來完成，主要步驟包括：(A) 遺失值處理、(B) 異常值處理、(C) 標準化、(D) 資料拆分及 (E) 特徵工程等。各項操作重點說明如下：

A. 遺失值處理

在機器學習資料前處理階段，需要處理原始水質資料中的遺失值（Missing Value），處理方式通常有兩種：直接刪除或進行填補。雖然直接刪除較為快速，但可能會對模型建構的完整性產生隱患，考慮到模型訓練需要連續分佈的資料，實作上可選擇以填補的方式保持水質資料的完整性，方式包括使用常見值填補、平均值填補、模型預測填補、相似案例填補或使用特殊標記表示遺失值 [24]。本研究將以 NumPy 及 Pandas 對水質資料進行遺失值處理，首先使用 NumPy 的 numpy.nan 表示缺失值，接著以 Pandas 提供的 isnull() 函式尋找資料中是否存在缺失值，最後使用 fillna() 函式進行填補。

B. 異常值處理

除了遺失值的處理外，異常值的檢測與處理也是資料清理中重要的一環。本研究使用修剪法（Trimming）排除機器學習模型資料集中的異常極端值，於整體水質資料集中，刪除中位數 1% 及 99% 之極端值，提升模型的性能 [25]。在經過資料清理排除水質資料集中的異常值後，最後剩餘 60,714 組水質資料作為模型資料集。

C. 標準化

正規化（Normalization）將數值資料縮放到 [0, 1] 的範圍，可能會導致異常值造成的資料遺失，因此使用標準化（Standardization）更佳。標準化將輸入參數的均值（μ=0）和方差（σ=1）進行調整，使資料更符合常態分佈 [26]。本研究使用 Scikit-learn 的 StandardScaler 函式進行標準化，將水質資料集中的數值資料進行標準化處理。

D. 資料拆分

E. 特徵工程

2-2-3. 模型選擇與比較

2-2-4. 超參數調整

2-2-5. 改良與測試

2-2-6. 實現系統化

3. 執行進度
[內容]

4. 遭遇之困難與解決途徑
[內容]

(三) 完成之工作項目及成果。

1. 完成之工作項目
[內容]

2. 完成之研究成果
[內容]

3. 學術研究、國家發展及其他應用方面預期之貢獻
[內容]

References

[1] Shao, Z., Zhao, R., Yuan, S., Ding, M., & Wang, Y. (2022). Tracing the evolution of AI in the past decade and forecasting the emerging trends. Expert Systems with Applications, 209, 118221.

[2] Cowls, J. (2021). 'AI for Social Good': Whose Good and Who's Good? Introduction to the Special Issue on Artificial Intelligence for Social Good. Philosophy & Technology, 34(Suppl 1), 1–5.

[3] Aggarwal, K., Mijwil, M. M., Al-Mistarehi, A. H., Alomari, S., Gök, M., Alaabdin, A. M. Z., & Abdulrhman, S. H. (2022). Has the future started? The current growth of artificial intelligence, machine learning, and deep learning. Iraqi Journal for Computer Science and Mathematics, 3(1), 115-123.

[4] Zhou, Y., Wang, X., Li, W., Zhou, S., & Jiang, L. (2023). Water Quality Evaluation and Pollution Source Apportionment of Surface Water in a Major City in Southeast China Using Multi-Statistical Analyses and Machine Learning Models. International Journal of Environmental Research and Public Health, 20(1).

[5] Im, Y., Song, G., Lee, J., & Cho, M. (2022). Deep Learning Methods for Predicting Tap-Water Quality Time Series in South Korea. Water (20734441), 14(22), 3766.

[6] Imran, H., Al-Abdaly, N. M., Shamsa, M. H., Shatnawi, A., Ibrahim, M., & Ostrowski, K. A. (2022). Development of prediction model to predict the compressive strength of eco-friendly concrete using multivariate polynomial regression combined with stepwise method. Materials, 15(1), 317.

[7] Narayan, V., & Daniel, A. K. (2022). Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model.

[8] Zeini, H. A., Lwti, N. K., Imran, H., Henedy, S. N., Bernardo, L. F. A., & Al-Khafaji, Z. (2023). Prediction of the Bearing Capacity of Composite Grounds Made of Geogrid-Reinforced Sand over Encased Stone Columns Floating in Soft Soil Using a White-Box Machine Learning Model. Applied Sciences, 13(8), 5131.

[9] Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). New York, NY, United States.

[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.

[11] Berrar, D. (2019) Cross-Validation. In: Ranganathan, S., Gribskov, M., Nakai, K. and Christian Schönbach, C., Eds., Reference Module in Life Sciences Encyclopedia of Bioinformatics and Computational Biology, Vol. 1, Elsevier, Amsterdam, 542-545.

[12] McKinney W. (2022). Python for Data Analysis: Vol. Third edition. O' Reilly Media.

[13] Harnowo, A. (2022). Blending a MOOC course into a Business School' s Course to Introduce Python for Data Analytics. Business Education Innovation Journal, 14(2), 31–35.

[14] Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.

[15] Nishino, R. O. Y. U. D., & Loomis, S. H. C. (2017). Cupy: A numpy-compatible library for nvidia gpu calculations. 31st confernce on neural information processing systems, 151(7).

[16] McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.".

[17] Chen, D. Y. (2017). Pandas for everyone: Python data analysis. Addison-Wesley Professional.

[18] Aurélien Géron. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition. O' Reilly Media, Inc.

[19] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learn-ing in Python, Journal of Machine Learning Re-search, 12.

[20] Hunt, J. (2023). Introduction to Matplotlib. In Advanced Guide to Python 3 Programming, Cham: Springer International Publishing, 121-128.

[21] Bill Lubanovic. (2024). FastAPI. O' Reilly Media.

[22] Bansal, P., & Ouda, A. (2022, July). Study on integration of fastapi and machine learning for continuous authentication of behavioral biometrics. In 2022 International Symposium on Networks, Computers and Communications (ISNCC) (pp. 1-6). IEEE.

[23] Native, R. (2020). React native. línea]. Disponible en: https://reactnative. dev/.[Último acceso: 2 de noviembre 2019].

[24] Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. International journal of computer science, 1(2), 111-117.

[25] Burlig, F., Knittel, C., Rapson, D., Reguant, M., & Wolfram, C. (2020). Machine learning from schools about energy efficiency. Journal of the Association of Environmental and Resource Economists, 7(6), 1181-1217.

[26] Sharma, P., & Singh, J. (2018, September). Machine learning based effort estimation using standardization. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 716-720). Greater Noida, NCR New Delhi, India.