透過人工智慧提升對淡水挖子尾濕地藍碳之含量評估效率與成本降低

二、研究計畫內容：

1. 摘要

隨著2022年3月底國發會發布2050淨零碳放路徑，碳中和、碳權、自然碳匯等議題也持續升溫，如何將空氣中的二氧化碳儲存起來已經是一個相當重要的課題，自然碳匯也是國發會所提出的淨零排放12項戰略之1[2]，海洋生態系被認定是一個能夠儲存很多碳的地方，紅樹林在海岸地帶，容易經營與保育，因而成為最為重要的儲探目標區，初步評估，紅樹林的碳儲存能力是所有藍碳當中最高[8]。

本計畫針對在淡水挖子尾溼地的三個採樣點的土壤進行分析，將採集好的樣本，經過幾項前處理步驟以後，將其放入EA元素分析儀與FTIR光譜儀當中進行分析，並將結合4種AI模型：（1）基於規則和實例的回歸模型(Cubist)；（2）隨機森林(Random forest)；（3）極限學習機(Extreme Learning Machine, ELM) ；（4） 偏最小平方回歸(Partial Least Square Regression, PLS)，透過AI模型來學習FTIR數據與元素分析儀數據之間的關聯性，在這過程中嘗試不同的參數設定以找出最理想的分析結果，將模型訓練完以後，透過三種驗證方法：（1） (R Squared, R平方) ；（2）RMSE (Root Mean Square Error, 均方根誤差) ；（3）sMAPE (Symmetric Mean Absolute Percentage Error, 對稱平均絕對百分比誤差)，來判斷模型的好壞程度。接下來，根據好壞程度的依據與模型效率及效能的花費以進行比較，比較後選定一種最合理的模型。最終得到最理想的模型與參數設定，在對土壤分析上透過紅外光與AI來配合分析，能夠省去使用分析儀的時間成本與成本降低。

1. 研究動機與研究問題
   1. 研究動機與背景介紹

國發會於民國111年3月底發表了台灣2050淨零排放路徑規劃，宣示要在2050時達到淨零排放的目標，利用負碳技術成為一個很重要的碳中和方式，因此藍碳就是一個相當重要的方向。藍碳(blue carbon)這個概念首次正式由2009年的一次聯合國多機構(UNEP,FAO and IOC/UNESCO)聯席會議所提出，主旨為世界各國應加強研究、重視、保育與增進海岸生態系的吸碳 (absorbing carbon) 與封碳(sequestering/storing carbon)角色[8]。經過近十年的比較研究後，學者建議紅樹林的儲碳是國家層級最有效益的減碳途徑之自然解方[11]。因此，為了要增加紅樹林的儲碳，勢必需要進行碳含量的分析，針對這一個部分我希望能夠引入AI模型。此計畫所使用的土壤樣本採集與聯合國糧食及農業組織所使用的方法相同[6]，以下為我們目前的土壤採集與資料分析，對於土壤採樣的部分，土壤採樣地點位於淡水挖子尾濕地的三個採樣地點(圖1)，已由台灣大學地質科學系羅立助理教授的團隊完成，並以進行元素分析。逢甲大學團隊則進行FTIR光譜儀分析。三個採樣點中的一個採樣點位於沙灘區，另外兩個位於水鼻仔紅樹林區。紅樹林區的底質粒徑屬極細沙至細沙之間，適合水鼻仔生長[]。樣本採樣的工具，使用了兩種採樣工具，分別是俄羅斯岩芯器(Russian corer)與壓入式岩芯器(Push core)來進行採樣。根據目前已有的研究，透過尋找FTIR與元素分析儀之間的關聯性以分析碳含量，針對這部分我們會利用4種AI模型來協助進行，並且會從中挑選出最合理的模型，以此來加快碳分析的效率與降低成本。

一張含有 地圖 的圖片

自動產生的描述

圖 1. 淡水挖子尾採樣地點（台灣大學地質科學系羅立博士提供）

* 1. 研究問題

因為在取得碳含量前需要經過許多繁瑣的流程，同時花費一定的費用與時間成本，在透過元素分析儀分析樣本一個大約需花費16分鐘來完成。因此我們將透過分析元素分析數據與FTIR光譜數據的關聯性，並且透過AI來協助，在這個過程當中透過四個模型的比較與權重的條整，來找出最合理的模型與權重的設定。因此我的問題是，如何從這四種模型當中去找到最合理的模型與參數如何去設定能夠得到最好的結果，未來再透過掃描紅外光並且將資訊輸入模型進行分析，最終以達到加快碳分析的效率與成本的降低。

1. 文獻回顧與探討

根據參考文獻[5]，當中的研究內容包含了土壤性質的預測，基於所開發的基PLS(篇最小平方回歸)模型經由FTIR光譜儀數據來預測土壤當中的氮(N)、磷(P)、鉀(K)和其他有機質含量。

首先，他們將FTIR應用於土壤定量分析，獲取光譜數據，並使用PLS模型以確定土壤的氮(N)、磷(P)、鉀(K)和其他有機質含量，將土壤樣本分為兩組，一組為「訓練集」，進行Leave-One-Out交叉驗證與標定誤差(Calibration Error)、驗證誤差(Validation Error)、標準差(Standard Deviation)各種檢證，以此優化PLS因子；另一組則為「驗證集」，是用來驗證優化的PLS模型。針對數據的處理，他們使用了平滑濾波器(Savitzky–Golay濾波器)對光譜進行預處理，這種方法傾向於保留數據的特徵，例如：波峰等等。再來透過PLS對光譜數據進行分析。在PLS因子優化的部分，他們發現在PLS建模當中校準與驗證結果在很大程度上的取決於PLS因子的數量。最後的結果顯示土壤養分與許多土壤成分相關，因此他們採用了多元校準的方式來提取FTIR光譜當中的資訊。PLS回歸模型建立於光譜與土壤樣品中土壤養分含量的關聯性。以結果來看，其中最好的預測是針對有機碳，次要是氮(N)與磷(P)。

以上研究是針對土壤養分與土壤當中的有機碳含量透過PLS回歸模型來尋找關聯性。本研究與以上研究存有差異之處，因為我的研究方向是針對土壤當中的有機碳含量與FTIR光譜之間的關聯性。此外，本研究將會透過幾種模型來尋找關聯性並且預測，其中包含了PLS方法，根據學習以上期刊研究的內容，能夠評估評估並確立何種方法來進行光譜預處理與優化決定因數，以此來增進未來研究紅樹林藍碳儲存的順暢度。

1. 研究方法及步驟
   1. 土壤樣本處理與分析流程：

完成了採樣步驟後，先將樣本送到實驗室進行土壤方塊烘乾與土壤切片冷乾，接著將烘乾完的土壤進行切片並且磨成粉末，最後將粉末狀之土壤樣本放入EA元素分析儀與FTIR光譜儀來進行分析。

1. EA元素分析儀
2. 元素分析儀原理：採用高溫動態燃燒法，當儀器運作時，燃燒管定溫為1100~1150，透過重力原理垂直投入燃燒管，再以適當量的氧氣與錫之助燃下，樣品瞬間燃燒溫度將會達到1800，使得樣品能夠完全燃燒，再將燃燒過後所生成的氣體於還原管當中與銅粉反應，以轉變成等氣體，再通過特殊氣體將各元素分離，最後再經由熱傳導檢測器(TCD)進行檢測，最後則可測得碳、氫、 氮、硫之重量百分比。
3. 經由分析儀所得出之數據：將以處理好的土壤樣本放入分析儀後得出了三個採樣地點的ROC (Recalcitrant Organic Carbon, 惰性有機碳)、LOC (Labile Organic Carbon, 活性有機碳)、TOC (Total Organic Carbon, 總有機碳)、TIC (Total Inorganic Carbon, 總無機碳)、TN (Total Nitrogen, 總氮)、TC (Total Carbon, 總碳)之百分比含量，並且將其對應到每一筆數據的深度。目前的數據量為，地點A共有41筆、地點B共有50筆、地點C共有32筆，一共有123筆數據。

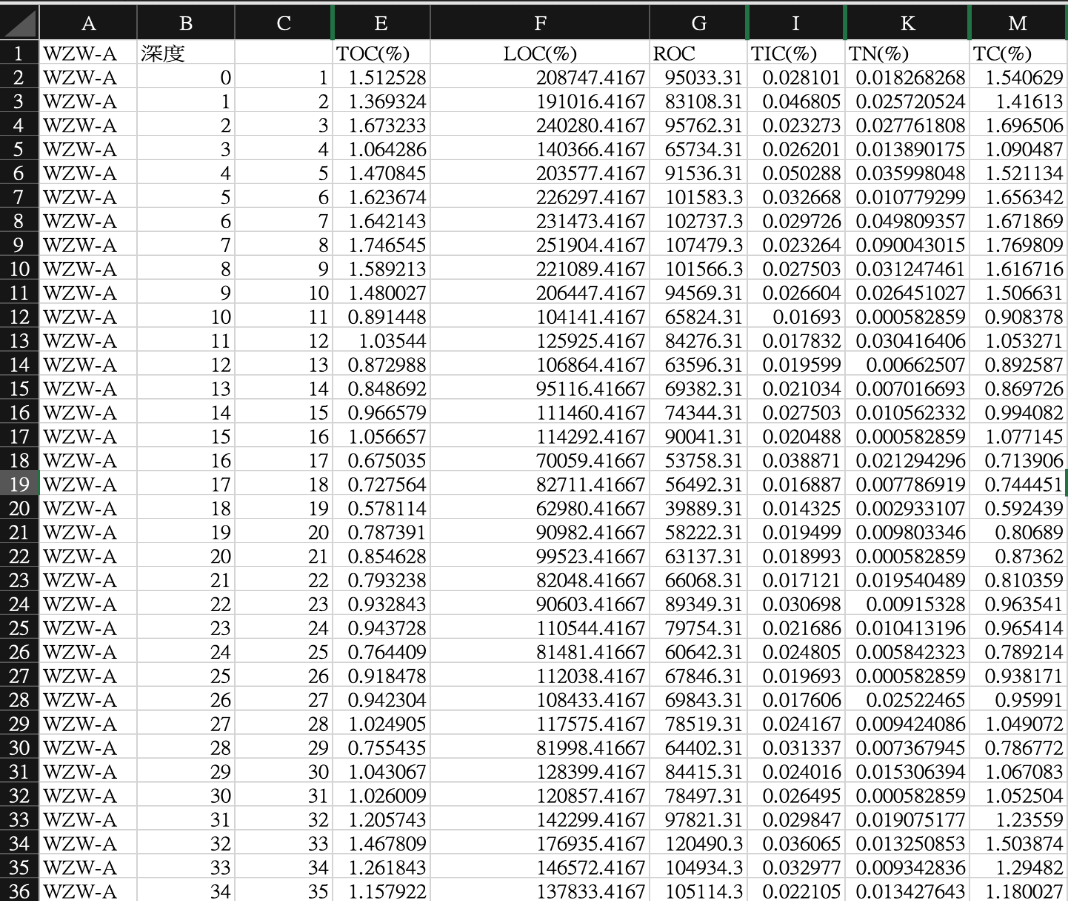


圖 2. 採樣地點A之元素分析數據。

1. FTIR (Fourier-Transform Infrared Spectroscopy, 傅立葉轉換紅外光譜)
2. 土壤FTIR分析：土壤樣本經過FTIR光譜儀能夠產出波數(Wavenumber)與穿透率(Transmittance)之數據。目前所具有三個採樣地點的數據資料量為，地點Ａ共有41筆、地點B共有50筆、地點C共有32筆，一共有123筆數據。
3. 數據初步分析：首先，我先針對FTIR資料進行初步分析，三個地點分別繪製同一地點不同深度的波形-穿透率之圖形以找出較特別的特徵。

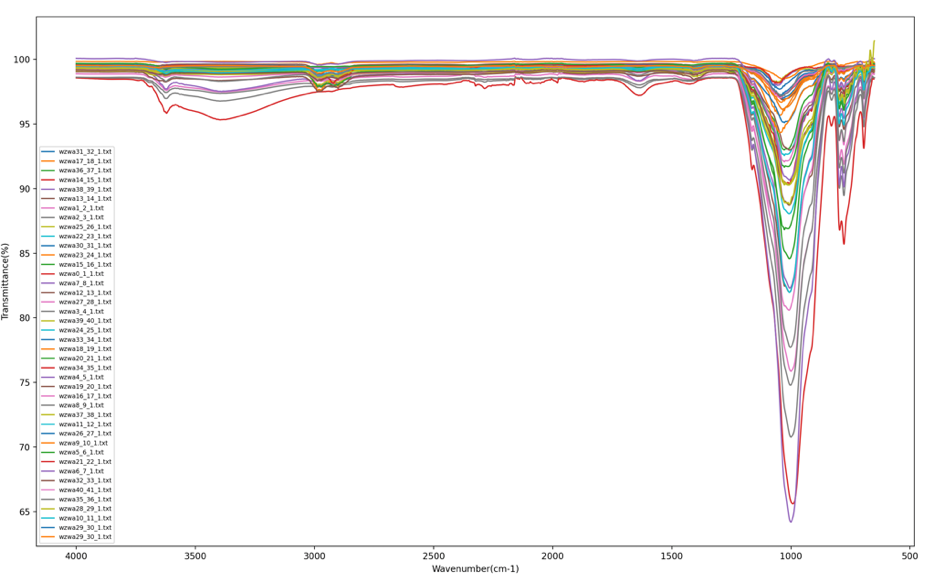


圖 3. 地點A之波數-穿透率圖形。

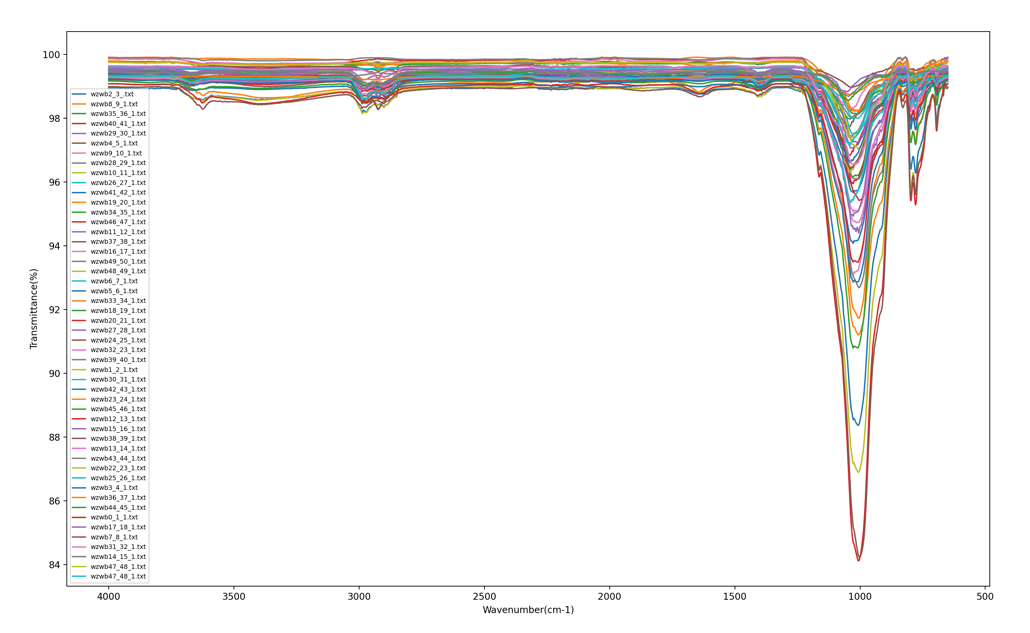


圖 4. 地點B之波數-穿透率圖形。

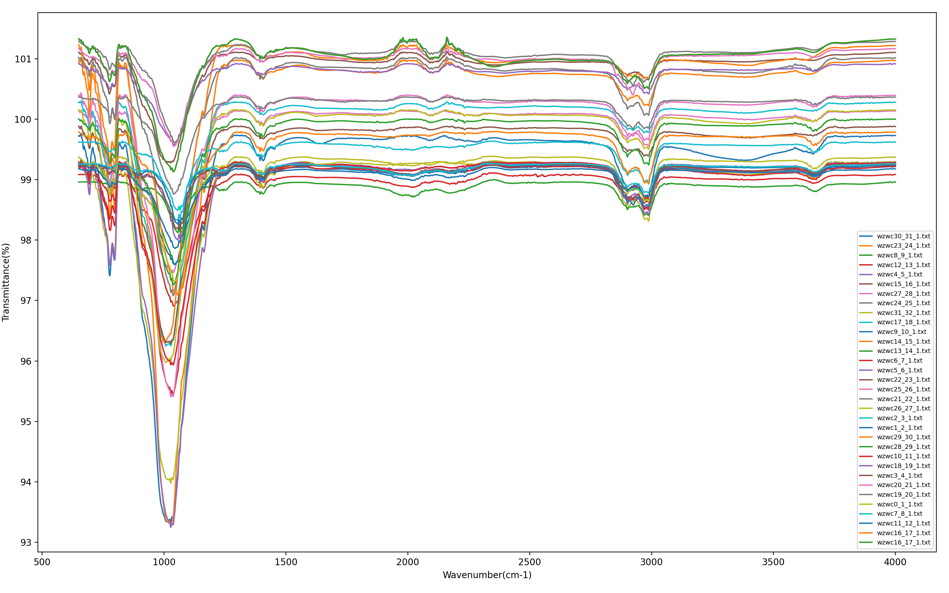


圖 5. 地點C之波數-穿透率圖形。

由上面三個圖形可以看出Ａ地點(沙灘)與Ｂ、Ｃ地點(紅樹林)土壤樣本的FTIR波譜特性有明顯的差異，暗示有機碳含量可由光譜特徵表現出來。

* 1. 土壤數據分析：

當我們數據準備完成後，將其進行數據分析，分析主要分兩個方向。第一點，透過波數與穿透率之數據對應到FTIR光譜波段所代表的土壤有機成份，再將資料進行相互關聯的標記。第二點，元素分析儀所得出的數據，針對數據內容進行初步分析，再透過與FTIR數據所對應到的資訊，並且透過機器學習來學習兩種(元素與光譜)數據之間的關聯性。

如下圖6，將元素分析儀數據裡面所含的項目數據與FTIR的數據進行比對，以尋找它們之間的關聯性，並且將這些關聯性透過AI來進行分析與學習。

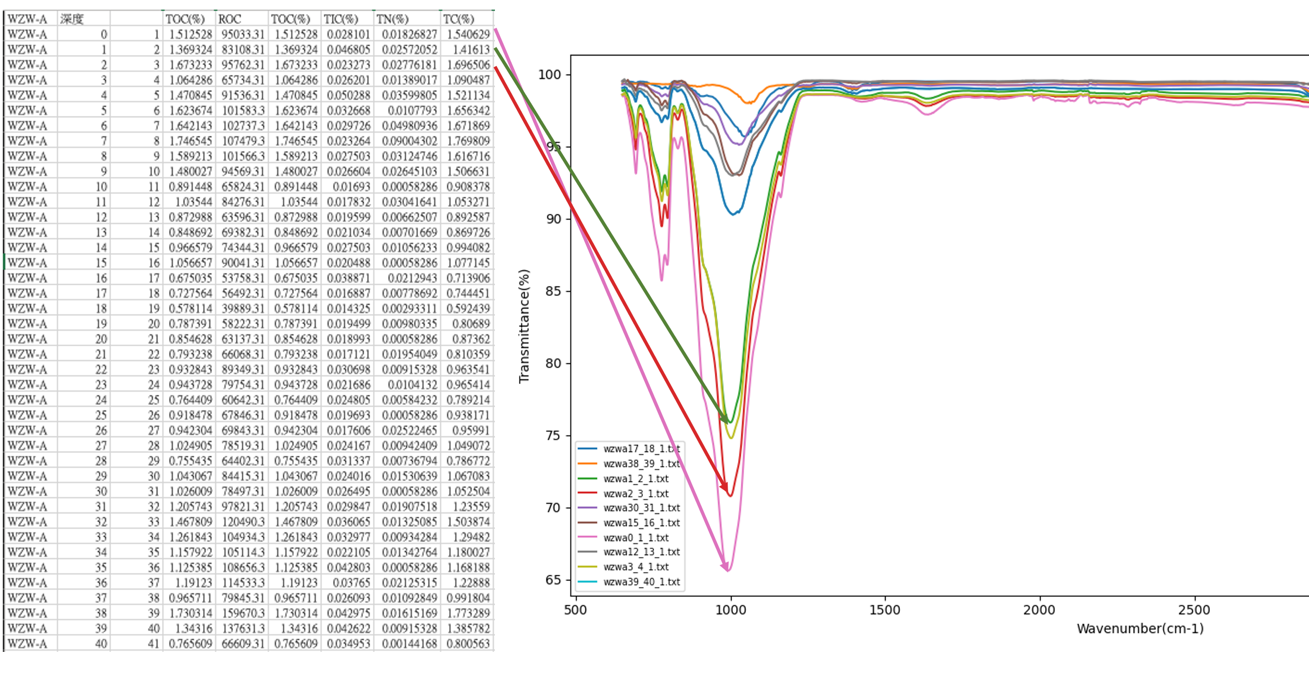


圖 6. FTIR與 元素分析數據之關聯性表示圖

* 1. 機器學習於FTIR數據和元素分析儀數據之分析：

人工智慧下包含了機器學習與深度學習，在機器學習裡面又分為監督式學習(Supervised Learning)、非監督式學習(Unsupervised Learning)以及強化學習(Reinforcement Learning)。以監督式學習來講，應用範圍非常廣泛，因為它擁有明確的學習目標，也就是數據中的目標值(Target Value)又稱標籤(Label)，期望能藉由預測(Prediction)或估計(Estimation)目標值。目標值有兩種類型，一種為連續數值、另一種為離散的類別，前者應用被徵為回歸，後者被稱為分類。非監督式學習，僅透過尋找特徵，也就代表它沒有目標的學習目標。主要應用為分群(Clustering)與維度縮減(Dimension Reduction)。強化學習，所做的並非為預測，而是建立一個連續決策(Continuous Decision)的代理人(Agent)，因此要最佳化決策行動(Action)，過程當中模型會根據代理人與環境(Environment)互動後所得到的報酬(Reward)與狀態(State)轉移的數據。模型會根據報酬的好壞自行逐步修正，以得到最佳的決策[1]。

我的計畫將會採用監督式的模型來進行訓練。使用監督式的原因，因為我們所要尋找的目標是碳含量，透過元素分析儀與FTIR的數據關聯性，透過回歸的方式以此來尋找我們的目標值。

1. AI模型
2. 基於規則和實例的回歸模型 (Cubist)

Cubist是由John Ross Quinlan[3]所開發的一種回歸演算法，主要適用於生成基於規則的預測模型。它使用了集成學習的想法，將多個回歸模型組合在一起，模型會透過對訓練集當中的資料尋找他們之間的關聯性，以建立一組回歸模型。如圖9所示，此圖裡面所要預測的目標為outcome，Cubist會去學習數據當中的其他項目與目標的關聯性並且生成出方程式。以圖7規則1為例，他的規則是假如petal\_length這個項目的數值小於等於4.7，就將各項目的數值帶入模型生成出的outcome預測方程式[9]。

* 優點：Cubist可以透過增加更多的模型來增加預測的準確度，這些模型可以針對先前的錯誤做修正。
* 缺點：可能會有過度擬和的情形發生，對於參數選擇需要適當地去調整。訓練所需時間較長。

選用Cubist的主要原因是因為此模型是去尋找數據之間的關聯性，與本計畫類型相同，透過此模型來去尋找FTIR與元素分析儀數據之間的關聯性。

一張含有 文字 的圖片

自動產生的描述

圖 7. Cubist輸出示意圖。本圖取自參考資料[12].Patrick Aselin 2021

1. 隨機森林 (Random Forest)

* 模型介紹：

隨機森林[4]是一種基於決策樹分類器的集成式學習演算法，由 L. Breiman 在 2001 年提出。隨機森林演算法中的分類器為CART (Classification And Regression Tree)，使用Bagging (又稱Bootstrap Aggregating) 演算法進行組合學習，並在CART樹進行生長時隨機選取變數以進行分類。CART演算法概念是使用遞迴二元分割(Recursive Binary Splitting)來歸納與分析大量複雜變數的資料集。

* 樹生成方法：

第一步，從訓練集當中抽取n筆資料(這n筆資料能夠重複抽取)。以我們目前數據為例，假設有123筆，從中抽取50筆數據，這50筆資料可能會有重複的數據。第二步，從這些資料當中挑選i個特徵當作決策因子，代表每一棵樹都只能看到部分的特徵。最後，重複以上步驟k次並產生k棵樹。當建立多棵決策樹後，將它們的預測結果結合起來。若為分類問題，則採用多數決的方法來整合預測結果，若為回歸問題，則採用平均法來整合預測結果。

* 優點：

第一點：因為隨機森林是基於CART演算法，故可以處理類別資料與連續資料。

第二點：透過集合多組決策樹的預測結果來提升準確度。

第三點：較不容易過度擬和。

第四點：能夠接受高維度的特徵資料

* 缺點：

第一點：運算時需要大量的資源，以儲存每棵樹的資訊。

第二點：因隨機森林是將多棵決策樹組和起來，因此無法對單一樹作解釋。

1. 極限學習機 (Extreme Learning Machine, ELM) 介紹資料之間的關係

* 模型介紹：

極限學習機[7]是一種單隱藏層前潰神經網路(SLFN)學習演算法，這種演算法只需要設置網路的隱藏層節點數量與激活函數。執行過程中不需要調整網路的輸入權重，在輸入層到隱藏層的權重是透過隨機產生，然後再使用最小平方法來確定輸出層的權重。ELM結構主要由兩層網路所組成，第一層先隨機初始化然後固定，第二層則是進行訓練。

* 訓練流程：

第一步：先輸入訓練集

，激活函數與隱藏層數量

第二步：模型隨機生成隱藏層權重與偏誤()

第三步：計算隱藏層的輸出矩陣

第四步：使用最小平方法來確定輸出的權重

* 優點：訓練速度快，ELM演算法可以在短時間內訓練完成且精準度高
* 缺點：解釋性較低，可能會出現過度擬和的情況

1. 偏最小平方回歸(Partial Least Square Regression, PLS)

* 介紹：

偏最小平方回歸是一種基於協方差的統計學方法。針對PLS在模型建立上，與多元線性回歸、主成分分析及其他重要的分析技術有關。PLS是一種用來建構預測模型的統計方法[10]。

1. AI模型之準確度驗證

針對本計畫所使用到的模型，為了驗證模型的好壞程度與回歸方程式的合理性，我將採用以下3種方法[1]：R平方(R Squared, ) ；RMSE (Root Mean Square Error, 均方根誤差) ；sMAPE (Symmetric Mean Absolute Percentage Error, 對稱平均絕對百分比誤差)。以下式子當中，定義實際值為y、預測值為、平均數為、樣本數為n。

1. R平方 (R Squared)：

R平方又稱判定係數(Coefficient Of Determination)，是用來評估回歸模型的表現，代表X解釋Y的能力，X能夠解釋Y變動的比例。針對本計畫，將透過R平方來檢驗模型所建立出的線性回歸式的可性程度。選用R平方的原因，因為本計畫所使用的模型均為回歸模型，R平方就是一種衡量回歸模型表現的指標。

* R平方公式：透過1減去總平方和分之殘差平方和。式(1)中的SSR(Sum Of Square Regression)為可解釋變異(Explained Variation)是預測值與平均值相差的平方和。SST(Sum Of Square Total)為整體變異，是實際值y與平均值相差的平方和。SSE(Sum Of Square Error)為不可解釋變異(Unexplained Variation)，將實際值y與預測值相差的平方和，如公式(1)所示。

|  |  |
| --- | --- |
|  | ( 1 ) |

* 殘差平方和(Sum Of Square Error, SSE)：透過真實數值 - 預測值即為殘差，為了計算方便，累加同時取平方。因此殘差平方和越大，也就代表模型解釋力越低，如公式(2)所示。

|  |  |
| --- | --- |
|  | ( 2 ) |

* 總平方和(Sum Of Square Total, SST)：實際值y與平均值相差的平方和，如公式(3)所示。

|  |  |
| --- | --- |
|  | ( 3 ) |

1. RMSE (Root Mean Square Error, 均方根誤差)：

RMSE是一種用於評估回歸模型的好壞損失函數，RMSE主要就是拿MSE去取根號，最後所得結果解釋起來會較直觀。MSE也就是拿預測的值與真實數據相減，所得的相減值平方過後並加總，最後再除以總數量平均，因此當MSE值越小，說明預測模型具有較好的精準度，如公式(4)所示。

* RMSE公式：

|  |  |
| --- | --- |
|  | ( 4 ) |

1. sMAPE (Symmetric Mean Absolute Percentage Error, 對稱平均絕對百分比誤差)：

MAPE是一種用於評估回歸模型的損失函數，透過計算預測值與真實數值之間的百分比誤差再取平均，將誤差轉換為百分比，主要對於解釋起來較直觀，也不需要去注意單位的問題，但MAPE存在三個缺點，第一點，分母為0時無法相除，第二點，若預測值誤差很大，指標可能會超過100%，第三點，實際值與預測值互換時，指標會產生不對稱的結果，如公式(5)所示。

* MAPE公式：

|  |  |
| --- | --- |
|  | ( 5 ) |
|  |  |

因此我將選用sMAPE，它是將MAPE的分母的分母進行修正為了解決不對稱的問題。整體來說，MAPE對於同樣的實際值平衡較好，sMAPE對於不同樣的實際值平衡較好，因此sMAPE對整體的誤差評估較為均衡，如公式(6)所示。

* sMAPE公式：

|  |  |
| --- | --- |
|  | ( 6 ) |

1. 預期結果
2. 模型與參數調整產出：經過模型的好壞比較與驗證過後，將選定最為合理的模型為本計畫的主要AI模。並且在過程當中進行模型的設定與參數與權重的調整，最終輸出最為合理的設定。
3. 時間成本減少：完成本計畫之研究以後，對於時間上可以省去分析的時間，以元素分析儀來看，能夠減少時間為n個樣本乘上16分鐘。
4. 花費減少：完成本計畫後，能夠減少樣本經由FTIR分析與元素分析儀的花費。
5. 投稿論文或結果發表：完成本計畫後，我打算向將本計畫的結果與模型、參數的設定等等的結果發表至期刊或論文，提供給對於這領域有相關研究或計畫的人使用。
6. 參考文獻
7. 李家岩, & 洪佑鑫 . (2022). 製造數據科學：邁向智慧製造與數位決策. 前程文化. <https://www.books.com.tw/products/0010922085?loc=P_0020_004>
8. 臺灣 2050 淨零排放 路徑及策略總說明. (2022). 國家發展委員會. <https://ws.ndc.gov.tw/Download.ashx?u=LzAwMS9hZG1pbmlzdHJhdG9yLzEwL3JlbGZpbGUvMC8xNTA0MC8yZTZhZTA0Mi0wYjUyLTQ4OTAtOGY5NC1hYjk5MzgzNWZlZTIucGRm&n=6Ie654GjMjA1MOa3qOmbtuaOkuaUvui3r%2bW%2bkeWPiuetlueVpee4veiqquaYji5wZGY%3d&icon=..pdf>
9. An Overview of Cubist. (n.d.). RuleQuest Research. <https://www.rulequest.com/cubist-win.html>
10. brownlee, J. (2016). Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch. Machine Learning Mastery. <https://books.google.com.tw/books?id=n--oDwAAQBAJ&hl=zh-TW&source=gbs_navlinks_s>
11. Du, C., Zhou, J., Wang, H., Chen, X., Zhang, J., & Zhu, A. (2009). Determination of Soil Properties Using Fourier Transform Mid-Infrared Photoacoustic Spectroscopy. Vibrational Spectroscopy, 49(1), 32–37. <https://www.sciencedirect.com/science/article/abs/pii/S0924203108000805?casa_token=ADrFU01xYY0AAAAA:wxdySxSA5p8EajEthY6e5zAG6qwTb34h2Lx-6dNhDooFzGbDOvHJFZsoHpXWh7jnmb7jR711V21B>
12. Fao/gsp. (2020). Soil Testing Methods Manual. FAO. <https://www.fao.org/documents/card/en/c/CA2796EN>
13. Huang, G., Zhu, Q., & Siew, C. Kheong. (2004). Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks. Proceedings of International Joint Conference on Neural Networks (IJCNN2004), 2. <https://www.researchgate.net/publication/4116697_Extreme_learning_machine_A_new_learning_scheme_of_feedforward_neural_networks>
14. Nelleman, C.; Corcoran, E.; Duarte, C.M.; Valdés, L.; DeYoung, C.; Foseca, L.; Grimsditch, G. (Eds.)　Blue Carbon: A Rapid Response Assessment; United Nations Environmental Programme and GRID-Arendal: Arendal, Norway, 2009.

<https://digitallibrary.un.org/record/673428>

1. Patrick Aselin. 2021. Cubist.

<https://github.com/pjaselin/Cubist.(2021)>.

1. PARTIAL LEAST SQUARES REGRESSION (PLS). (n.d.). XLSTAT. <https://www.sciencedirect.com/topics/nursing-and-health-professions/partial-least-squares-regression>
2. Taillardat, P., Friess, D. A., & Lupascu, M. (2018). Mangrove blue carbon strategies for climate change mitigation are most effective at the national scale. Biology Letters, 14(10), 20180251.

<https://doi.org/10.1098/rsbl.2018.0251>

1. 需要指導教授指導內容
2. 協助指導教授
3. 土壤分析/藍碳分析：

由於土壤與藍碳都並非是我有接觸過的內容，對於一些詳細內容的理解與土壤分析數據裡面數值所代表的意義，將會請教魏國彥教授(前國立台灣大學地質系教授)。

1. 數據預處理與光譜數據特徵觀察：

針對光譜數據，如何去進行標準化以及去異質化，對於如何去驗證該種方法對於數據所產生的好、壞影響，該透過哪種方法去呈現並且進行調整，以及在光譜數據當中應該如何去評估當中的特徵。這部分將會請教梁辰睿教授(材料科學與工程學系助理教授)。

1. 指導教授
2. AI模型選擇與概念理解：

針對光譜數據，透過什麼方法能夠抓取特徵，對於數據之間的關聯性可以透過什麼樣的方式來標記，這方面將會請教指導教授，針對程式撰寫標記方法與如何分析關聯性。

1. 光譜數據與元素分析數據特徵抓取與關聯性分析：

對於不同的數據類型與型態都會對應到不一樣的AI模型，AI模型的選擇上，以我的研究計畫的資料類型來看，對於要使用何種模型較為合理並且對於AI模型的運作方法，需要指導教授進行相關概念的指導。

1. AI模型驗證與調整的建議：

針對本計畫所選定的AI模型表現程度，需要透過驗證的方法來決定，在驗證模型好、壞以後，對於模型互相的比較也包含了的運作效率與效能消耗程度的比較。對於模型權重的設定、參數的調整、模型是否過度擬和等等，AI模型的驗證與如何調整AI模型需要指導教授的指導。