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（以10頁為限）：

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，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的。

本計畫針對紅樹林的土壤進行分析，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，將分析完的數據資料尋找他們之間的關聯性，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)，並且透過三種驗證方法：（1）R平方 (R squared) ；（2）RMSE (Root Mean Square Error, 均方根誤差) ；（3）MAPE (Mean Absolute Percentage Error, 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，最後選定一種最適合的模型。本計畫透過結合AI模型以提升整體的效率並且降低成本。

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。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。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模型來增加效率與成本降低。

* 1. 研究問題

現階段若要分析碳含量則需要透過元素分析儀來取得土壤當中的碳儲量，但進行分析前經過繁瑣的流程：(1)到各個地點進行土壤採集；(2) 將土壤樣本進行烘乾；(3) 將土壤磨成粉(一公分取一公克磨粉) ； (4) 最後將處理好的土壤放入元素分析儀進行分析。經過一連串的步驟，不只需要花費大量的時間與高額的成本。

1. 文獻回顧與探討

在參考文獻[1]當中，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. 土壤樣本採樣：
2. 土壤採樣地點介紹：

土壤採樣地點位於淡水挖子尾濕地的三個採樣地點，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。



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. 採樣工具介紹與採樣流程：

採用兩種採樣工具，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。

1. 俄羅斯岩芯器(Russian corer)：此工具為一根半圓形鐵管與長條鐵片，能夠採集將近一米的樣本，單次取樣最長為50cm。選擇此器材是因為能夠獲得較長且連續的樣本，係將它插入土壤裡後將其旋轉抽出，即可取得未經翻轉或混合的完整岩芯的樣本，樣本抽出後會把樣本放在包有鋁箔的半圓形鐵罐，以方便後續的處理。

一張含有 地面, 個人, 室外, 泥土 的圖片

自動產生的描述

圖 2. 使用俄羅斯岩芯器(Russian corer)採樣過程。

1. 壓入式岩芯器(Push core)：此工具為一根中空的圓形鐵管，有上下蓋子，與”Russian corer”不同之處在於深度的不同，壓入式岩芯器大約只能取得10 cm的岩芯，且後續處理土壤的方法不同。採樣流程係將鐵管插入土壤當中，用手掌壓著上面的孔後，再將其慢慢拔出以取得岩芯。



圖 3. 壓入式岩芯器(Push core)。

* 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。

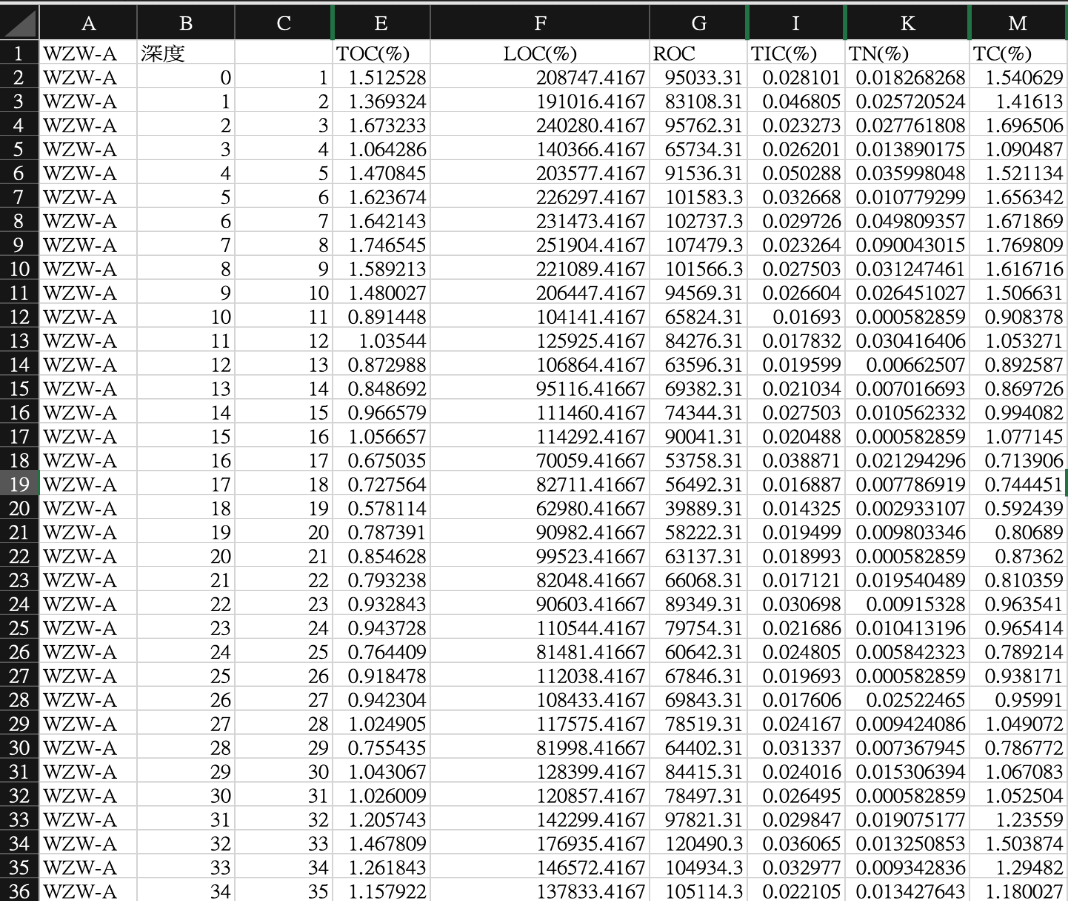


圖 4. 採樣地點A之元素分析數據。

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。

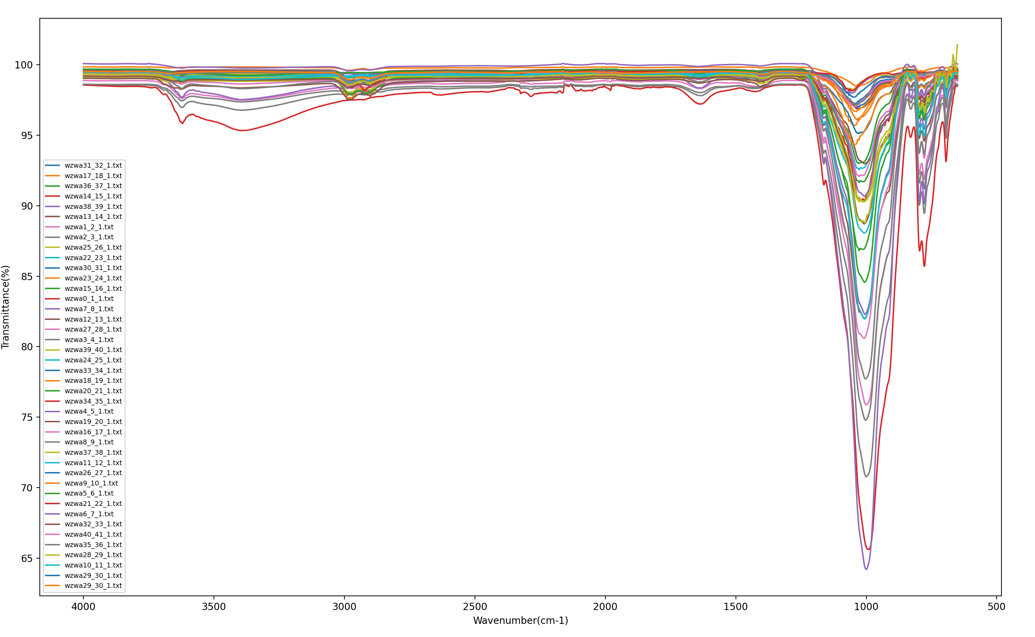


圖 5. 地點A之波數-穿透率圖形。

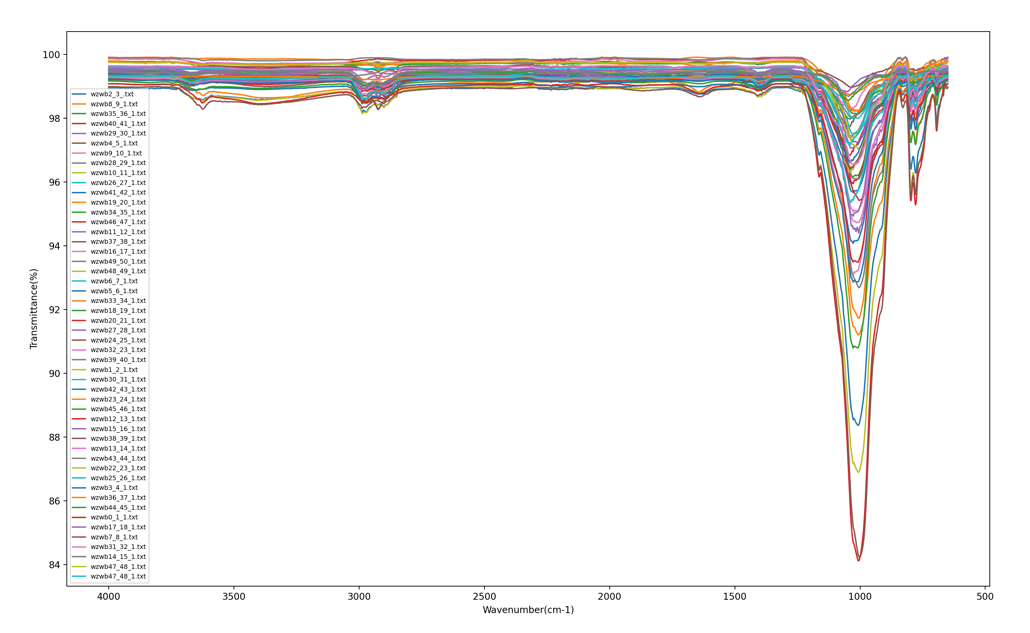


圖 6. 地點B之波數-穿透率圖形。

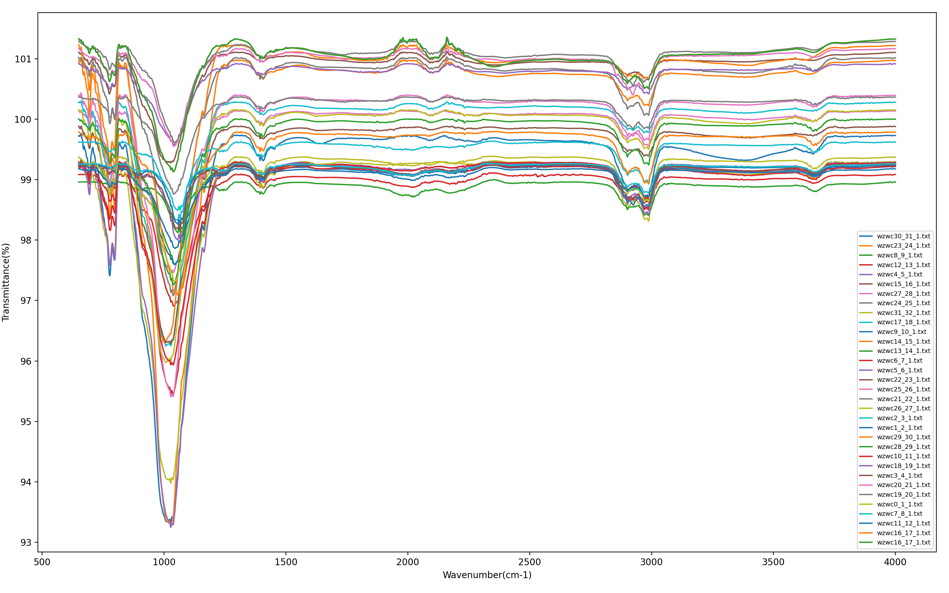


圖 7. 地點C之波數-穿透率圖形。

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. 土壤數據之數據分析：

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相對應的官能基，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。

* 1. 機器學習於FTIR數據和元素分析儀數據之分析：

(補充與我們計畫的關聯性)

1. AI模型
2. 基於規則和實例的回歸模型 (Cubist)

Cubist是由John Ross Quinlan [ ]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。

* 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。

一張含有 文字, 水鳥, 鳥, 植物 的圖片

自動產生的描述

圖 8. Cubist輸出示意圖(取自)

1. 隨機森林 (Random Forest)

* 模型介紹：

隨機森林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演算法概念是使用二元分割（原文）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。

1. AI模型之準確度驗證
2. 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。

* R平方公式：透過1減去總平方和分之殘差平方和
* 殘差平方和 (residual sum of squares)：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。
* 總平方和：以真實值減去平均值取平方

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。

* RMSE公式：

1. MAPE (Mean Absolute Percentage Error, 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。

* MAPE公式：

1. 預期結果

(列點)

透過本計畫研究，能夠改善以往傳統的土壤有機碳分析方式，利用分析FTIR光譜分析結果建立與土壤屬性的關聯性，來改善傳統透過元素分析儀的時間與成本問題，並且結合AI模型，透過不同種類的模型來互相比較，以得出最為合適的一款AI模型。

透過以上的預期結果，使得台灣在紅樹林的儲碳增進上能夠更加的有效率且減少成本的花費，對於台灣2050的減排路線，根據此種方法，來幫助提升台灣在儲碳上的效率。

1. 參考文獻
2. 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>
3. Vestergaard , R., Vasava, H. B., Aspinall , D., Chen , S., Gillespie , A., Adamchuk , V., & Biswas, A. (2021). Evaluation of Optimized Preprocessing and Modeling Algorithms for Prediction of Soil Properties Using VIS-NIR Spectroscopy. Sensors, 21(20). [https://www.mdpi.com/1424-8220/21/20/6745#](https://www.mdpi.com/1424-8220/21/20/6745)
4. An Overview of Cubist. (n.d.). RuleQuest Research. <https://www.rulequest.com/cubist-win.html>
5. Heuer, J. (2019, November 18). Cubist Models in R: Balancing Interpretability and Predictive Power. ODSC. <https://opendatascience.com/cubist-models-in-r-balancing-interpretability-and-predictive-power/>
6. Wang, J. (2018, March 26). 隨機森林 (Random Forest). Rstudio. <https://rstudio-pubs-static.s3.amazonaws.com/378052_30d987a09ea54b6db5aa1e82f5dce6bf.html>
7. [Day 14] 多棵決策樹更厲害：隨機森林 (Random Forest). (2021, September 26). IT邦幫忙. <https://ithelp.ithome.com.tw/articles/10272586?sc=hot>
8. 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>
9. Extreme Learning Machine. (n.d.). Wikipedia. （不能用）<https://en.wikipedia.org/wiki/Extreme_learning_machine>
10. PARTIAL LEAST SQUARES REGRESSION (PLS). (n.d.). XLSTAT. <https://www.sciencedirect.com/topics/nursing-and-health-professions/partial-least-squares-regression>
11. 需要指導教授指導內容
12. 數據預處理

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。這方面需要老師提供一些建議，透過什麼方法能夠將數據做最好的處理。

1. 光譜數據與元素分析數據特徵

針對光譜數據，應該透過什麼樣的方式評估該抓取什麼特徵，並且透過什麼方法能夠抓取特徵，另外，對於數據之間的關聯性可以透過什麼樣的方式來標記。這方面需要有相關經驗的老師來指導，針對程式撰寫標記方法，與如何尋找關聯性。

1. AI模型選擇與概念理解

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，需要老師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模型需要老師的指導。