Python與機器學習

農作物的水質數據集

水質參數(for馬鈴薯)

M132040016 邱榮材
M132040025 邱威誠

目錄

一、引言

(一) 研究背景與目的:

1. 水質對於作物生長的重要性已被廣泛研究，尤其對於需水量大的馬鈴薯，其灌溉水質會直接影響產量與品質。
2. 本研究利用機器學習模型分析水質指標對馬鈴薯生長的影響，並嘗試預測水質是否適合作物生長。
3. 本研究的目的是通過模型比較，尋找準確率更高且泛化能力更強的模型，為農業灌溉決策提供支持。

(二) 研究問題:

1. 哪些水質指標對馬鈴薯生長影響最大？
2. 透過機器學習模型，如何有效預測水質的適合性？

二、資料概述

(一) 資料來源

此資料集從Kaggle上「Water Quality Dataset for Crop」取得，內容包括對馬鈴薯最佳生長重要的水質指標，資料及涵蓋PH值、水質硬度、溶解固體量等不同參數。

(二) 變數說明與資料前處理

1. 變數說明**:**

1.PH:

- 定義:衡量水溶液酸性或鹼性的指標，範圍從 0（酸性）到 14（鹼性），PH = 7 為中性。
- 影響:通常馬鈴薯適合生長在微酸性至中性的土壤中，水質需求在5.5~6.5。如果水質的酸鹼度過酸或過鹼，可能會抑制養分吸收，導致生長不良或產量下降。

2.硬度（Hardness）:

- 定義:水質硬度主要由水中鈣離子和鎂離子含量決定，分為「軟水」與「硬水」。

- 影響：適量的硬水（鈣和鎂）對馬鈴薯生長有益，因為鈣是細胞壁形成的重要元素，鎂是葉綠素的核心成分。過高硬度（極硬水）可能導致土壤結構改變，影響根部透氣性和水分吸收。

3.溶解固體（Solids）：
- 定義：總溶解固體又稱（Total Dissolved Solids, TDS) 包括水中可溶的礦物質、鹽類、有機物等。
- 影響：適量的溶解固體可提供養分（例如鈣、鎂、硫等），有助於馬鈴薯生長。高濃度 **TDS** 可能導致鹽害，抑制水分進入根部，嚴重時會造成生長停滯甚至植株死亡。

4.氯胺（Chloramines）：
- 定義：氯胺是水消毒過程中加入氯氣後形成的化合物，用於殺菌，對植物具有毒性。
- 影響：低濃度 氯胺對土壤和作物影響有限，但長期灌溉高氯胺水可能影響土壤微生物群落，進而影響馬鈴薯根部養分吸收。灌溉水中的氯胺濃度應低於 0.5 mg/L。

5.硫酸鹽（Sulfate）：
- 定義：水中硫酸根離子濃度，硫元素是植物生長必需的次要營養元素之一。
- 影響：硫酸鹽是馬鈴薯生長所需的必須養分，可促進蛋白質與酶的合成，增強抗逆性。過量硫酸鹽 可能導致鹽害，影響植株的水分平衡和根部發育。

6.溶解有機碳（DOC）：
- 定義：水中有機物質分解後的主要成分，以碳原子含量當作判斷的指標。
- 影響：適量的有機碳可提升土壤有機質，改善土壤結構，增強保水保肥能力，有利於馬鈴薯根系發育。過量有機碳 可能引發土壤厭氧環境，抑制根系呼吸，導致生長不良。

7.三鹵甲烷（THMs）：
- 定義：在水消毒過程中，氯與有機物反應生成的副產物，具有潛在毒性。
- 影響：三鹵甲烷可能通過灌溉進入土壤，對馬鈴薯植株產生潛在毒害。長期暴露可能抑制根部生長，降低作物產量，並對土壤生態造成負面影響。

8.濁度（Turbidity）：
- 定義：水中懸浮顆粒物（如泥沙、有機物）所造成的渾濁程度。
- 影響：濁度高的水質可能攜帶病菌與污染物，影響土壤健康，增加病害風險。過多的懸浮物可能覆蓋土壤表層，阻礙空氣與水分的滲透，影響根系生長。

9.Check:

- 定義:判斷水質是否適合馬鈴薯生長，類別為0和1（0:不適合，1:適合）。

欄位說明總結:對於馬鈴薯生長，水質的各項指標至關重要

1. 適當的 PH 值與硫酸鹽濃度 可促進營養吸收與代謝。
2. 水硬度與溶解固體 提供必需礦物質，但過量會造成鹽害。
3. 有機碳與濁度 需控制在合理範圍內，避免土壤污染和病害。
4. 氯胺與三鹵甲烷 需特別留意，過量會對作物造成毒害。

2. 資料探索與前處理**:**

1. 缺失值:

   在PH值有491筆缺失值，硫酸鹽有781筆，三鹵甲烷有162筆。
   因為缺失值佔比較大，所以選擇使用各列的中位數去填補。

2. 資料分佈:

   圖2-1為所有數值變數的分佈圖，可以分布為鐘型分布，並不需要做變數轉換。

3. 觀察不同群組差別:

   對check做分類，將PH值分成三群:<5.5為低PH值 5.5~7.5為中PH值，>7.5為高PH值。分別對這兩種分群做分析，發現各變數在不同群組並無區別。

4. 新增交互項與相關係數:

   圖2-2為各變數的相關係數矩陣，根據圖形可以發現各欄線性關係是較低的，所以新增增加特徵”硬度x硫酸鹽”、”溶解固體x硫酸鹽”、”PHx硬度”，圖2-3即增加後的相關係數矩陣，但並沒有有相關性高的類別。

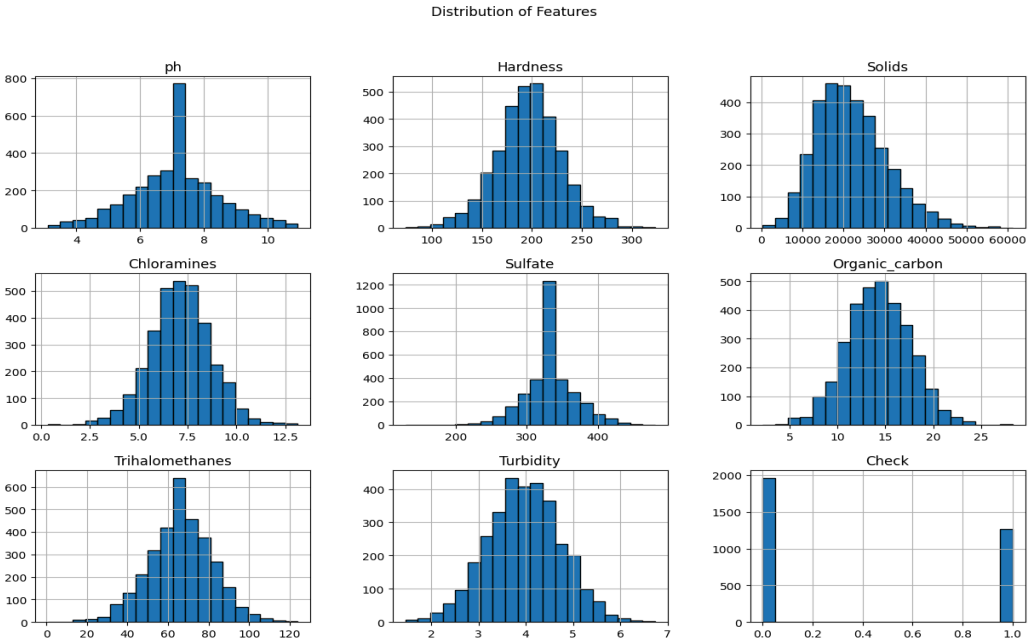

圖2-1 各變數的分布圖

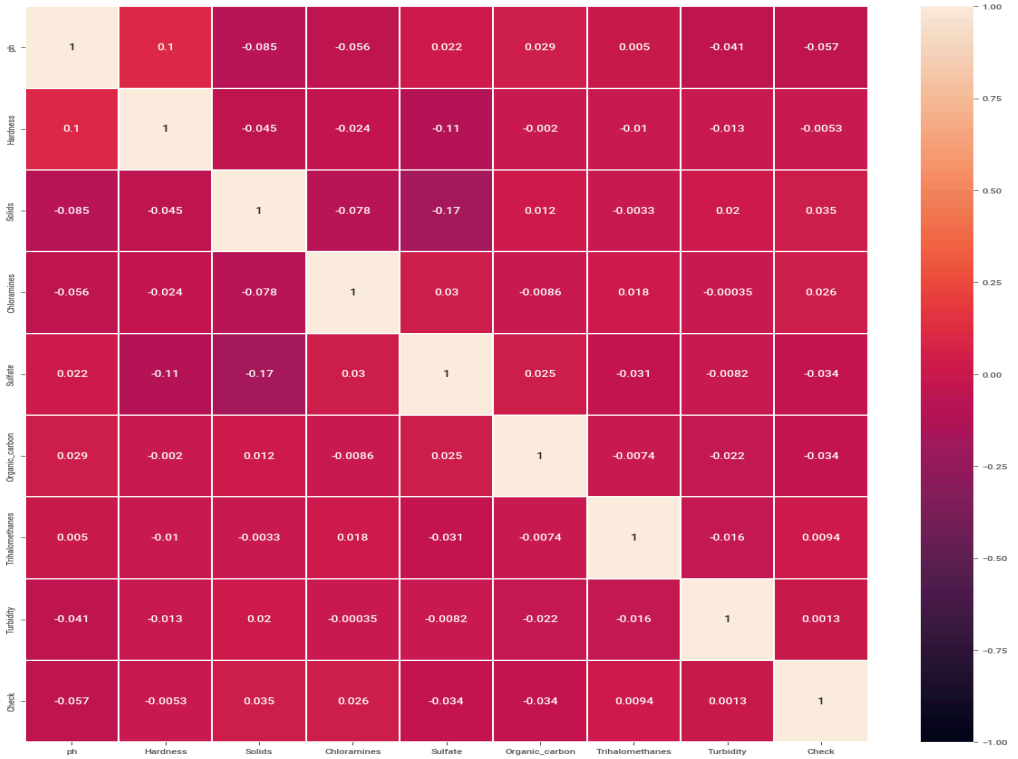

圖2-2相關係數熱力圖

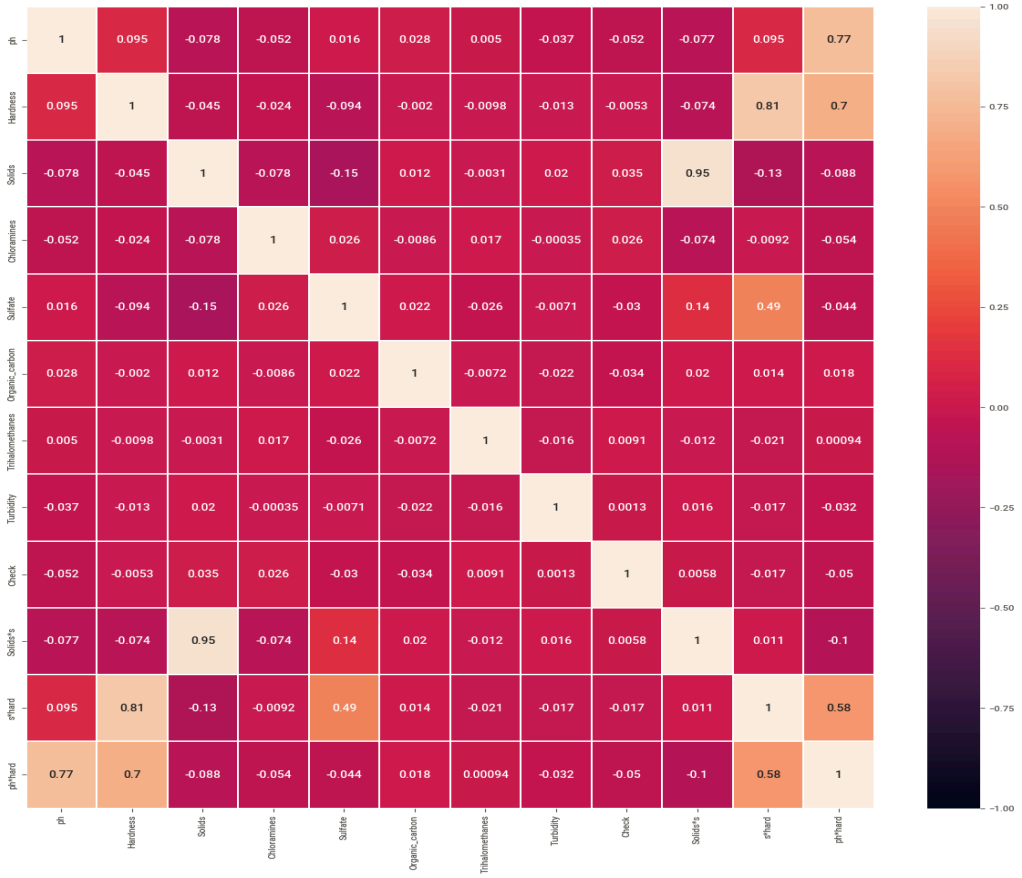

圖2-3 增加交互項的熱力圖

三、模型建立與比較

(一)模型選擇

為了預測目標變數 **Check**（水質檢查結果），我們採用了多種機器學習模型，包括：

- **Decision Tree Regressor**
- **ExtraTree Regressor**
- **CatBoost Regressor**
- **Random Forest Regressor**
- **KNeighborsClassifier**

目的是比較不同模型的預測準確度，並透過參數調整來優化模型表現。

下列是各模型的介紹:

- **Decision Tree**：

  是一種基於樹形結構的監督式學習模型，適用於分類和迴歸任務。它透過遞歸式地對特徵進行分割，構建一棵「決策樹」來預測目標值。

- **ExtraTree**：

  與 Decision Tree 類似，但在進行特徵分割時引入了更高的隨機性。與傳統決策樹相比，ExtraTree 不僅隨機選擇特徵，還隨機選擇分割點，從而增加模型的多樣性。

- **Random Forest**：

  是一種集成學習方法，由多棵隨機生成的決策樹組成。每棵樹在訓練時使用不同的隨機子集，最終通過「投票」或「平均」來獲得預測結果。

- **CatBoost**：

  是一種基於梯度提升（Gradient Boosting）的決策樹模型，特別針對類別特徵（Categorical Features）進行優化，具有高效能和高準確度。

- **K-Nearest Neighbors**：

  是一種基於「鄰近性」的非參數監督式學習方法。預測時，模型會尋找目標樣本周圍最近的 K 個鄰居，並根據鄰居的標籤進行分類或迴歸。(以下簡稱 KNN)

(二)初始模型表現

在一開始的實驗中，所有模型均使用 **Python** 預設參數 進行訓練與測試，模型表現如下：

- DecisionTreeRegressor 準確度: 0.5734
- ExtraTreeRegressor 準確度: 0.5517
- CatBoostRegressor 準確度: 0.6553
- RandomForestRegressor 準確度: 0.6445
- KNeighborsClassifier 準確度: 0.5363

觀察到初始結果，模型的準確度普遍不高，這顯示在預設參數設定下，模型未能充分學習數據中的規律。我們進一步對各模型的超參數進行調整，目的是提升準確度並防止過度擬合。

(三)模型參數調整

3.1 Decision Tree Regressor

在初始結果的基礎上，我們調整了 Decision Tree 的主要參數如下：

- **max_depth=8**：限制樹的深度，避免過度擬合，增加泛化能力。
- **min_samples_split=10**：最小分裂樣本數從 2 提高到 10，減少不必要的分支。
- **min_samples_leaf=5**：最小葉子樣本數從 1 提高到 5，保證葉節點有更多數據點。

3.2 ExtraTree Regressor

ExtraTree 與 Decision Tree 類似，我們的參數調整與Decision Tree 相同：

- **max_depth=8**：同樣限制樹的深度以控制模型複雜度。
- **min_samples_split=10**：提高分裂所需的最小樣本數。
- **min_samples_leaf=5**：增加葉子節點的最小樣本數，避免過度擬合。

3.3 Random Forest Regressor

我們針對樹的數量及單棵樹的複雜度進行調整：

- **n_estimators=200**：將樹的數量從預設的 100 提高到 200，增加模型穩定性。
- **max_depth=10**：限制樹的最大深度，防止過度擬合。
- **min_samples_split=10**：提高最小分裂的樣本數。
- **min_samples_leaf=4**：增加葉子節點的最小樣本數。

3.4 KNN

KNeighbors 模型的表現受距離度量和權重方式的影響，我們調整了以下參數：

- **weights = 'distance'**：將權重從「均等」改為「距離加權」，使得較近的樣本對預測結果有更大貢獻。
- 標準化數據：在模型訓練前，對數據進行標準化處理，確保所有特徵具有相同的尺度。

3.5 CatBoost Regressor

對於CatBoost，我們調整了以下參數：

- **iterations=2000**：增加樹的數量，讓模型有更多的迭代學習機會。
- **learning_rate=0.03**：新增學習率，保證模型訓練的穩定性。
- **depth=8**：增加樹的深度，提升模型擬合能力。
- **l2_leaf_reg=5**：添加正則化項，減少過度擬合風險。

- **bagging_temperature=0.8**：引入適度的隨機性，提升泛化能力。
- **early_stopping_rounds=50**：當模型在 50 次迭代內未改善時提前停止，避免浪費計算資源並防止過度擬合。

四、模型結果分析與特徵選擇比較:

(一)各模型的指標分析:

表4-1為原始模型的正確率、precision、recall score、和F1 score。可以了解到CatBoost的正確率是最高的，但也無法超過0.7，且recall score較低。相對於原始模型，表4-2為調整模型後的各項指標，也是CatBoost最高，調整過的模型正確率除了隨機森林都有提高，recal score除了KNN有降低。而KNN是調整後表現最好的。表4-3則是各模型調整後的參數減去調整前的參數的表格，正數代表調整後的指標較好。負數代表調後前較好

表4-1原始模型的各項指標

	DicisionTree	ExtraTree	CatBoost	RandomForest	KNN
正確率	0.5858	0.5966	0.6538	0.6491	0.5363
precision	0.4958	0.5095	0.6641	0.6211	0.4055
recall score	0.4398	0.5000	0.3195	0.3759	0.2744
F1 score	0.4661	0.5047	0.4314	0.4684	0.3273

表4-2更改參數模型的各項指標

	DicisionTree	ExtraTree	CatBoost	RandomForest	KNN
正確率	0.6429	0.6213	0.6630	0.6321	0.6506
precision	0.6131	0.6263	0.6864	0.625	0.6349
recall score	0.3158	0.2331	0.3045	0.282	0.4511
F1 score	0.4169	0.3397	0.4219	0.3886	0.5275

表4-3 調整前後的模型對比

	DicisionTree	ExtraTree	CatBoost	RandomFore	KNN

				st	
正確率比較	0.0571	0.0247	0.0092	-0.0170	0.1143
precision比較	0.1173	0.1168	0.0223	0.0039	0.2294
recall比較	-0.1240	-0.2669	-0.0150	-0.0939	0.1767
F1比較	-0.0492	-0.1650	-0.0095	-0.0798	0.2002

(二)各模型的特徵重要性:

1. Decision Tree Regressor:

調整前除了氯胺以外無明顯特徵，調整後氯胺、PH值、硫酸鹽較其他有明顯特徵。而SHAP解釋分布較平均，特翁重要性並不明顯。

2. ExtraTree Regressor

調整前除了PH值以外無明顯特徵，調整後PH值、硫酸、硬度較其他有明顯特徵。而SHAP解釋分布較平均，特翁重要性並不明顯。

3. Random Forest Regressor

調整前特徵不明顯，調整後更強調硫酸鹽、硫酸鹽*硬度、ph、和 氯胺這幾個特徵的重要性。

4. CatBoost Regressor

調整前後都以硫酸鹽和PH值為重要特徵。

5. KNN

調整前後都以硫酸鹽和溶解固體量、硬度的交互項有影響。

五、結論:

結論來講，雖然調整過後的模型效果正確率有明顯上升，但整體來講KNN和CatBoost的模型是最好的。即使是最好的正確率也才沒有大於0.7，根據Kaggle上的資料有人做到0.7左右，代表我們這裡的模型參數需要再細緻一點。

六、組員分工表:

邱威誠:簡報、模型、圖表、簡報、書面
邱榮材:大綱、找資料、EDA、上台報告、簡報、書面

七、參考資料:

https://www.kaggle.com/datasets/iamtapendu/crop-production-data-india