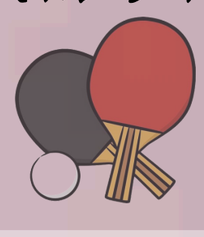
AI人工智慧運動科學數據分析:
集成學習模型可解釋性探討-應用於世界頂尖桌球選手之技戰術分析

國立成功大學　統計學系
申請人：張凱華

計畫編號：114-2813-C-006-022-M
指導教授：鄭順林



摘要

本研究旨在研究AI人工智慧方法中集成學習(Ensemble Learning)模型的可解釋性，應用運動科學數據分析於世界頂尖女子桌球選手之技戰術分析。研究的對象聚焦在現今女子桌球世界排名第一的孫穎莎選手，本研究透過邱宏達-鄭順林教授共同開發的桌球技戰術編碼系統，將比賽中的技術、戰術過程進行數據化，模型方面則選用極限梯度提升樹演算法(XGBoost)與隨機森林(Random Forest)兩種集成學習模型，再搭配 SHAP 方法(SHapley Additive exPlanations)來進行模型可解釋性探討，了解比賽中的技戰術使用對得失分的影響。

研究動機與研究問題

1. 研究動機

(1) AI人工智慧方法中集成學習模型(XGBoost、Random Forest、LightGBM)在複雜與高維數據上表現優異，但缺乏可解釋性。
(2) 在運動分析中，若僅有精確預測卻無法說明依據，教練與選手難以獲得具體技戰術指導，使應用價值受限。

2. 研究問題

(1) 如何提升集成學習模型於運動科學數據分析中的桌球比賽的可解釋性?
(2) 能否透過技術動作、球速、旋轉、落點與方向長短等變數，建構得失分模型並進行有效的技戰術分析?

選擇使用之模型

1. XGBoost

極限梯度提升樹(eXtreme Gradient Boosting，XGBoost)是一種基於梯度提升樹（Gradient Boosting Trees）演變的強大機器學習演算法，其基本概念是透過每棵新建的樹 $f_t(x_i)$ 的預測結果，修正前一棵樹的預測誤差(圖一)。模型主要學習目標是最小化以下目標函數 $L(t)$，為了方便最小化目標函數，XGBoost使用二階泰勒展開近似其中之損失函數 $l(y_i, \hat{y}_i)$，這是XGBoost與傳統梯度提升樹(GBDT)的主要差異之一。

$$L(t) = \sum_{i=1}^{n} \left[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

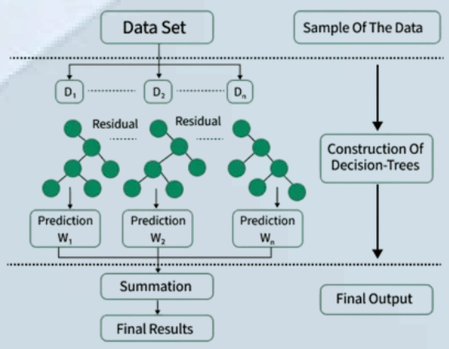
最終模型的預測結果即是先前所有樹的預測加總。

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F$$

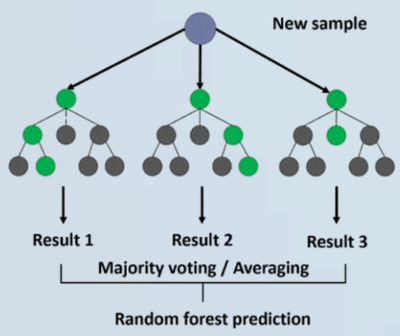
2. Random Forest

隨機森林的基本概念是從訓練集進行還原隨機抽樣，形成多個不同子訓練集，再隨機訓練出多個決策樹模型(圖二)，決策樹分裂準則大多根據基尼指數，每次分裂時，隨機森林會選擇能夠最小化基尼指數的特徵來進行分裂，這樣可以使每顆樹的節點內類別更乾淨，其中 $h(x;\theta_k)$ 代表第k棵樹的分類器，基於隨機向量 θ_k 構建，並將它們的結果進行投票或平均，來提高模型的穩健性和準確性。

$$RF = \{h(x, \theta_k), k = 1, 2, \ldots, K\}$$



圖一 : XGBoost模型流程示意圖



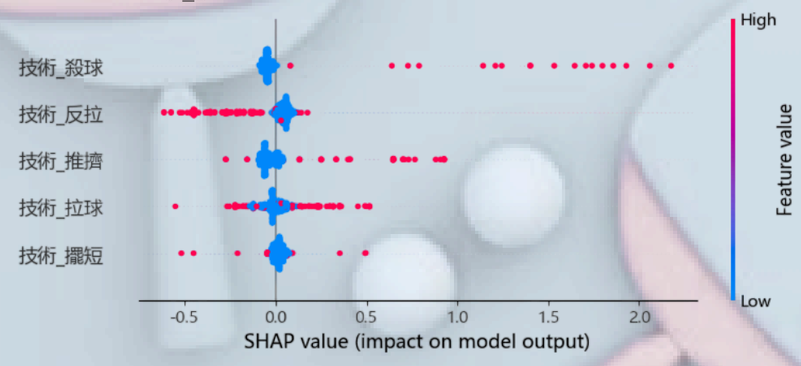
圖二 : 隨機森林模型流程示意圖

模型解釋之工具

Shapley Additive exPlanations(SHAP)

SHAP方法是本研究中解釋模型的重要工具，一種基於博弈論的模型解釋方法，其核心理念源於 Shapley 值，是博弈論大師 Lloyd Stowell Shapley 基於合作賽局理論提出，$\phi_i(f, x)$ 表示在考慮所有排列組合下的情況下，特徵 i 的 Shapley 值，SHAP方法主要將Shapley值這概念引入機器學習(圖三)，藉以公平的量化每個特徵對模型預測的貢獻。

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus \{i\}) \right]$$



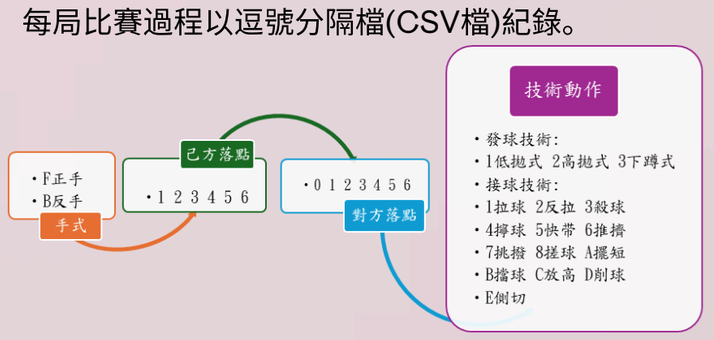
圖三 : 孫穎莎擊球技術之SHAP Summary Plot

研究方法及步驟

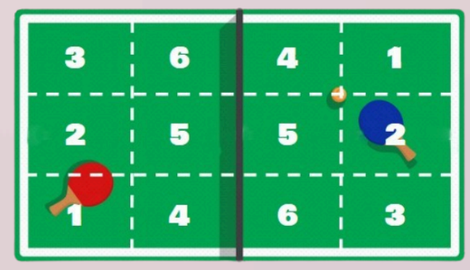
本研究使用 python 作為實作的主要語言

1. 資料蒐集

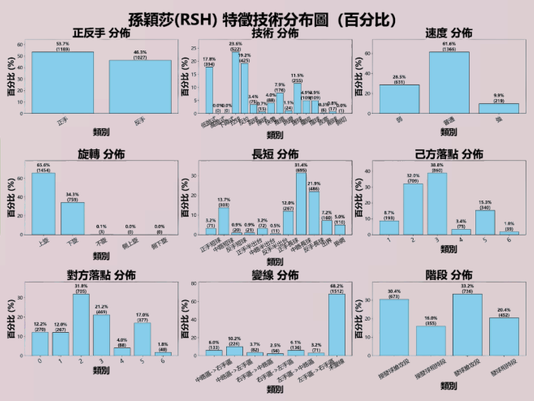
此資料蒐集根據國立成功大學邱宏達教授和鄭順林教授共同定義之編碼系統(圖四)，紀錄技術動作、速度、旋轉、落點、方向長短，將比賽影像轉為自定義編碼，此編碼系統將每局比賽過程以逗號分隔檔(CSV檔)紀錄。



圖四 : 邱宏達-鄭順林編碼系統單板紀錄流程圖



圖五 : 雙方球桌劃分示意圖



圖六 : 孫穎莎特徵分布圖

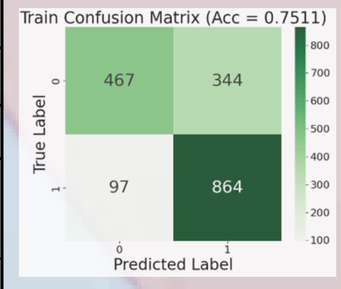
2. 特徵定義與模型擬合

本研究會將目標選手孫穎莎之所有擊球板資料進行模型擬合，再為其對手之所有擊球板資料進行模型擬合，意味著我們將為孫穎莎及其對手各自擬合極限梯度提升樹模型(XGBoost)和隨機森林(Random Forest)之得失分模型，再從雙方視角去解讀比賽，解釋變數選擇上(表一)，包含擊球位置、落點位置、技術動作、球速、旋轉和方向長短、階段、變線等多模態特徵(圖五及圖六)，而應變數則是該選手之得失分情形。

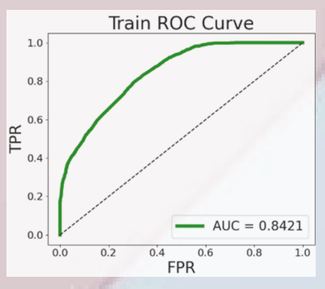
表一 : 解釋變數欄位與變數名稱對應表

括弧內為變數數量	變數名稱	球速 (3)	Speed_1(弱), Speed_2 (普通),Speed_3(快)
擊球手式 (2)	batting_hand (正手或反手)	旋轉 (5)	spin_Topspin(上旋),..., spin_SideBackspin(側下旋)
擊球位置 (6)	batting_position_1,..., batting_position_6	方向長短 (11)	Dire_forshort(正手短球),..., Dire_net(掛網)
落點位置 (7)	landing_position_1, ..., landing_position_6 landing_position_0	所處階段 (4)	serviceattack(發球搶攻)， receiveattack(接發球搶攻), servicestalemate(發球相持) receivestalemate(接發球相持)
技術動作 (3+13)	tech_ LowToss(低拋發球), ...,tech_sidewaypush(側切)	變線 (7)	linechange_l_to_m(左至中),..., linechange_no_change(未變線)

初步配適孫穎莎之XGBoost模型



圖七 : 模型混淆矩陣



圖八 : 模型ROC曲線

3.結果呈現方式及初步分析

(1) 不同模型分析角度
(i) 從孫穎莎模型出發：探討孫穎莎之特定技術動作或戰術選擇對得失分的影響。
(ii) 從對手模型出發：分析孫穎莎在面對某些技術或戰術時是否失分率較高等等。

(2) 視覺化圖表之呈現
(i) 直方圖 (Histogram)：呈現不同技術動作的出現比例或擊球落點。
(ii) 熱力圖 (Heatmap)：點分佈與戰術傾向（如攻擊反手區）。
(iii) SHAP Summary Plot：協助揭示影響戰術選擇與比賽結果的關鍵因素。

(3) 初步分析
(i) 以孫穎莎選手方之擊球板初步配適XGBoost 模型，如圖七及圖八。
(ii) SHAP 可解釋性分析(圖三)：使用「拉球技術」時，多數 SHAP 值為正，表示該技術有助於提高模型對得分的預測；相較之下，「放高技術」的 SHAP 值則高度集中於零點附近，且幾乎皆為藍點 (代表未使用該技術)，顯示目前的資料量不足以支持更深入的分析，仍需持續蒐集更多數據，使分析結果可以更具應用價值。

預期結果

討論AI人工智慧方法中極限梯度提升樹演算法(XGBoost)與隨機森林(Random Forest)。兩種皆以決策樹(Decision Tree)為基礎單元，卻具有不同運作機制的機器學習模型，並探討它們在運動科學數據分析中桌球技戰術分析的優劣勢，並瞭解哪些技戰術對比分的影響最為顯著，協助我國桌球好手鄭怡靜、黃怡樺、陳思羽、簡彤娟等選手面對世界頂尖好手時，能掌握對手各類技術動作及策略運用，並讓教練團隊能量身打造更具針對性的訓練計畫，在訓練中強化相應的攻守端能力，在面對世界排名前列的中國、日本、韓國等強敵時，能夠更有把握地應對。

重要參考文獻

Song, H., Li, Y., Zou, X., Hu, P., & Liu, T. (2023). Elite male table tennis matches diagnosis using SHAP and a hybrid LSTM–BPNN algorithm. Scientific Reports, 13(1), 11533.
Sun, M., Hsu, M., & Chen, H. (2022). 桌球選手廖振珽在男單比賽前五板之技戰術特徵分析. 興大體育學刊, 21, 25-34.