Supervised machine learning-based prediction for Flavours of Physics: Finding $\tau \to \mu\mu\mu$

A.P. Guan-Ju Peng
Institute of Data Science and Information Computing
National Chung Hsing University
Taichung, Taiwan
gjpeng@email.nchu.edu.tw

Hao-Wei Huang
Institute of Data Science and Information Computing
National Chung Hsing University
Taichung, Taiwan
nchuds110012hwh@smail.nchu.edu.tw

***Abstract*—標準模型 (The Standard Model，SM) 是 21 世紀的前沿物理之一，其目的是想探究宇宙萬物的組成與如何產生相互的作用力，在 2013 年實驗物理學家們終於找到標準模型最後一個理論預測的粒子 – 希格斯玻色子 (Higgs boson)，至此標準模型初步完備。然而，諸多在標準模型所無法解釋的現象一一出現，例如為何會出現微中子震盪 (neutrino oscillation，2015 年諾貝爾物理學獎) 的現象等問題，其中 2015 年 Kaggle 競賽之 Flavours of Physics: Finding $\tau \to \mu\mu\mu$ 就是其中標準模型所無法解釋的輕子風味破壞 (Lepton Flavour Violation) 現象，然而該衰變過程尚未確定，需要更長期且穩定的證據，本次競賽使用機器學習進行二分類預測，目的在於幫助實驗物理學家關注於感興趣的衰變過程 (例如 $\tau^- \to \mu^- + \mu^- + \mu^+$)，也就是說分類為 signal(感興趣事件) 與 backgroud(不感興趣事件，或是稱為 noise)，競賽同時要求該分類器必須同時通過真實數據與模擬數據的分布應該相似 (Agreement-test) 與分類依據應該與質量無關 (Correlation-test) 之要求。本專題將使用邏輯斯迴歸 (Logistic Regression，LR)、決策樹 (Decision Tree)、隨機森林 (Random Forest，RF)、AdaBoost、K-Nearest Neighbor 與近期流行的極限梯度提升方法 (eXtreme Gradient Boosting，XGBoost) 等機器學習方法架構分類器來評估與比較。實驗比較後，利用物理學公式新增特徵使用 XGBoost 方法 (不透過 PCA 降維) 表現最好，其 accuracy 約 86.5%、F1 score 約 89.2%、AUC(Area under the ROC Curve) 約 85.34，並且同時通過題目所要求的 Agreement、Correlation 檢驗!

***Index Terms*—Kaggle, The Standard Model, Lepton Flavour Violation (LVF), Machine Learning, Particle Physics**

I. 前言

A. 標準模型 *(The Standard Model)*

1900 年 4 月，英國物理學家 William Thomson 在英國皇家學會（The Royal Society）發表”Nineteenth century clouds over the dynamical theory of heat and light”[2]，說明當前物理幾乎完成只剩下修飾而已，但目前遇到兩個重大問題尚未解決，這誕生 20 世紀的量子力學 (quantum mechanics) 與相對論 (Theory of relativity) 的發展；在 21 世紀物理學家想探究宇宙究竟由什麼組成、到底如何作用於這世界提出不同的理論，其中標準模型 (The Standard Model) 提供一個有力的理論闡釋宇宙萬物組成以及作用力的來源，這需要量子力學與相對論的基礎才能建構。標準模型認為，物質由費米子 (fermion) 組成、作用力由玻色子 (boson) 傳遞，其中費米子分成夸克 (quark) 和輕子 (lepton)，各有三個世代 (generation)、每個世代有兩個粒子且粒子皆有反粒子；而玻色子根據作用力不同傳遞不同粒子，電磁力 (electromagnetic interaction) 由光子 (photons，γ) 傳遞、強作用力 (strong interaction) 由膠子 (gluons，g) 傳遞及弱作用力 (weak interaction) 由 W 玻色子與 Z 玻色子傳遞，但用數學對每個作用建模時會引用規範場論 (gauge theory)，這導致玻色子必須是無質量此與實驗矛盾 (W、Z 玻色子有質量)，故英國物理學家 Peter Ware Higgs 等 6 人分別解釋並提出後來的希格斯機制 (Higgs mechanism)，這賦予了 W、Z 玻色子質量伴隨希格斯玻色子出現，2013 年 3 月歐洲核子研究組織（CERN）使用大型強子對撞機（Large Hadron Collider，LHC）蒐集數據後公布終於找到標準模型下的最後一塊拼圖—號稱上帝粒子的希格斯玻色子，至此標準模型初步搭建完畢!

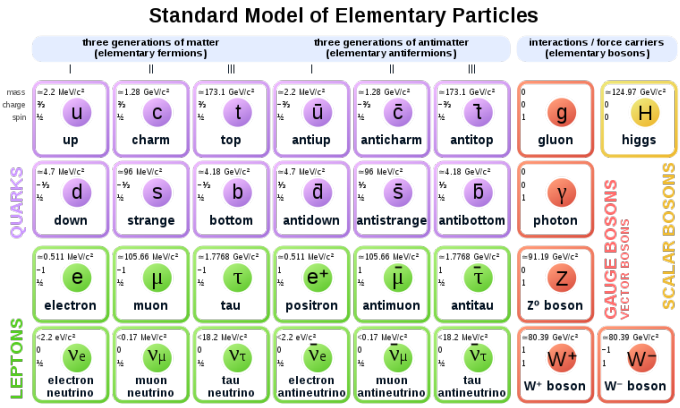


Fig. 1: 標準模型 (SM) 之架構

B. 輕子味荷破壞 *(Lepton Flavour Violation)*

標準模型的完成，引起物理學界的對於找尋夢寐已久的萬有理論（Theory of Everything）有了興趣，但同時該理論出現不少難以解釋的問題，例如費米子為什麼是三個世代 (2008 年諾貝爾物理學獎其中發現夸克至少三個世代) 而不是更多? 為什麼光子與膠子沒有質量? 為什麼微中子 (Neutrino) 具有質量? 上述這些問題引發物理學界的興趣，並稱之為「超越標準模型的物理學」(Physics beyond the Standard Model，BSM)，例如在輕子中就有部分實驗發現某些粒子衰變時會導致「輕子風味破壞」，即輕子數不守恆 [3]。L (輕子數) = 所有輕子 (e^-, μ^-, τ^-, ν_e, ν_μ, ν_τ)

TABLE I: 費米子與玻色子比較

比較	費米子	玻色子
目的	構成物質	傳遞作用力
分類	夸克、輕子	規範、純量
總數	48 (36 + 12)	13 (12 + 1)
遵循統計	Fermi-Dirac	Bose-Einstein
Pauli exclusion principle	遵守 (物質穩定)	不遵守 (物質穩定)
自旋 (spin) 數	半整數 ($\frac{1}{2}$, ⋯)	整數 (0,1, ⋯)
全同粒子	反對稱	正對稱
反粒子 (antiparticle)	皆有對應粒子	自身 ($W^{\pm}$ 互為反粒子)
靜止質量 (invariant mass)	有	部分沒有 (γ, g 沒有)
物理意義	任意兩顆不能同一量子態	任意數量可以同一量子態
舉例	電子	光子

的數量–所有反輕子 (e^+, μ^+, τ^+, $\bar{\nu_e}$, $\bar{\nu_\mu}$, $\bar{\nu_\tau}$) 的數量，在標準模型要求下輕子數應該守恆 (衰變前後的輕子數相同)，例如 $\mu^- \rightarrow\ e^- + \bar{\nu_e} + \nu_\mu$ 在衰變前的輕子數為 1、衰變後的輕子數 = 1 + (-1) + 1 = 1，故符合輕子數守恆；另一方面也可以定義 L_e (電輕子數) = 所有電輕子 (e^-, ν_e) 的數量–所有反電輕子 (e^+, $\bar{\nu_e}$) 的數量、L_μ (μ 輕子數) = 所有 μ 輕子 (μ^-, ν_μ) 的數量–所有反 μ 輕子 (μ^+, $\bar{\nu_\mu}$) 的數量、L_τ (τ 輕子數) = 所有 τ 輕子 (τ^-, ν_τ) 的數量–所有反 τ 輕子 (τ^+, $\bar{\nu_\tau}$) 的數量，它們應該同時滿足各自輕子家族數守恆 (conservation of lepton family number)，上述必須全部守恆才吻合標準模型要求。(非輕子或是非所屬輕子家族則為 0)。本次競賽的衰變 $\tau^- \rightarrow\ \mu^- + \mu^- + \mu^+$，L: 衰變前的輕子數為 1、衰變後的輕子數 = 1 + 1 + (-1) = 1 (守恆)、L_e: 衰變前的電輕子數為 0、衰變後的電輕子數 = 0 + 0 + 0 = 0 (守恆)、L_μ: 衰變前的 μ 輕子數為 0、衰變後的 μ 輕子數 = 1 + 1 + (-1) = 1 (不守恆) 及 L_τ: 衰變前的 τ 輕子數為 1、衰變後的 τ 輕子數 = 0 + 0 + 0 = 0 (不守恆)，因此該衰變過程若真實存在，則打破標準模型的要求! (迄今，實驗物理學家仍繼續驗證該過程是否存在)

C. 競賽意義 *(The Meaning of This Competition in SM)*

本次競賽主要是作為實驗物理學家的研究資料前的預處理 (data preprocessing)，因為數據基本上是數萬、甚至上百萬筆數據，裡面有些數據可能不是所關注的，例如本次競賽的 $\tau^- \rightarrow\ \mu^- + \mu^- + \mu^+$ 衰變，不可能總是發生，因此必須先將不重要的訊息 (粒子物理學稱之 background) 與需要關注訊息 (粒子物理學稱之 signal) 進行分類，也就是機器學習所謂的二分類問題，故本次競賽的本質就是做監督式學習下的二分類問題。

II. 研究方法

A. 資料說明 *(Dataset Description)*

原始共包含四份檔案: training.csv (訓練集)、test.csv (測試集)、check_agreement.csv (競賽限制 - 驗證模型是否一致) 與 check_correlation.csv (競賽限制 - 驗證模型與質量無關)，其中因為是監督式學習，因此資料會有標籤 (label) 欄位，其定義為”signal”(僅在 training.csv 和 check_agreement.csv 檔案有)。為了符合專題需求，test.csv 不作為驗證 (因為沒有標籤)，因此將 training.csv 隨機拆分 70% 作為新訓練集以作為模型建構、30% 作為新訓練集以作為驗證模型效能，此外為配合競賽要求，需注意兩項系統限制 [1]:

- Agreement-test (一致性檢定): 因為訓練資料包含真實數據跟模擬數據 (例如使用 Monte Carlo simulation)，因此系統希望在真實與模擬保持一致，而這份檔案提供另一種衰變資料: ${D^+}_S \rightarrow\ \pi^- + \phi\ (\rightarrow\ \mu^- + \mu^+)$，其與 $\tau^- \rightarrow\ \mu^- + \mu^- + \mu^+$ 的信號衰減之拓樸結構 (topology) 類似，同時也是 LVF 的可能證據之一 (尚未確認是否存在)，來驗證模型是否達到真實與模擬分布相似的限制，透過 Kolmogorov–Smirnov (KS) 檢定來評估樣本的差異性，其概念是利用累積分布函數 (cumulative distribution funnction，CDF) 來檢定真實與模擬分布是否相似，檢定量為:

$$KS = \max |F_{real} - F_{simulation}| \tag{1}$$

  其中，F_{real} 代表 *check_agreement.csv* 真實數據的 CDF、$F_{simulation}$ 代表 *check_agreement.csv* 模擬 (Monte Carlo simulation) 數據的 CDF，競賽要求 KS < 0.09。
- Correlation-test (相關性檢定): 實際上所有費米子 (構成物質的基本單位，例如 ${D^+}_S, \tau^-$ 等) 都擁有質量，但數據所蒐集到的質量只是預估值 (estimation) 因此使用該特徵很可能會誤導分類器的效能，故競賽要求分類器所分類的依據不應該與質量有關，其透過 Cramer–von Mises (CvM) 檢定來評估分類器是否與質量相關，其概念亦透過 CDF) 來檢定，檢定量為:

$$\begin{aligned} &CvM_{interval} = \int \left(F_{global} - F_{interval}\right)^2 dF_{global} \\ &CvM \leq CvM_{interval\ >interval} \end{aligned} \tag{2}$$

  其中，F_{global} 代表 check_correlation.csv 所有數據的 CDF、$F_{interval}$ 代表 check_correlation.csv 包含質量數據的 CDF，競賽要求 CvM < 0.002。

B. 分類問題方法 *(Method for Binary Classification)*

使用各種機器學習方式對數據搭建二分類模型，包含邏輯斯迴歸、支持向量機、隨機森林與當前在各式競賽與論文效能不錯的 XGBoost(例如 2014 年同時是 CERN 舉辦的 Kaggle 競賽中大放異彩) 等機器學習方法進行模型評估與比較。主要比較下列機器學習方法，並同時加入降維 (dimension reduction)、透過公式擴增其他物理學特徵嘗試是否能提升精準及通過 Agreement-test、Correlation-test。

a) 邏輯斯迴歸 (Logistic Regression, LR): 為監督式學習方法 (supervised learning)，最早用於統計模型上利用 sigmoid function 來模擬靠近 0 時機率應有變化並且可以微分，其輸出值 [0, 1] 故可用於二分類問題。此方法利用 sklearn package 的 linear_model.LogisticRegression 進行實驗。

b) 決策樹 (Decision Tree, DT): 為監督式學習方法，利用 if-else 架構將特徵進行分支，本質上是對特徵空間進行分割，其評估分支方法常見有 Information gain、Gain ratio、Gini index 等。此方法利用 sklearn package 的 tree.DecisionTreeClassifier 進行實驗。

c) 隨機森林 (Random Forest, RF): 為監督式學習方法，總體來說是決策樹的集合，每個 (棵) 樹建立是從訓練資料中重複抽樣 (resampling)，故其為 Bagging 與決策樹之結合。此方法利用 sklearn package 的 ensemble.RandomForestClassifier 進行實驗。

d) AdaBoost: 為監督式學習方法，是綜合多個分類器的方法，又本質是 Boosting 方法所以不同分類器皆是用同一份資料集並逐步減少 bias。此方法利用 sklearn package 的 ensemble.AdaBoostClassifier 進行實驗。(每個弱分類器皆為 DecisionTreeClassifier initialized 且固定最大深度 max_depth 為 1)

e) K-Nearest Neighbor, KNN: 為監督式學習方法，藉由設定 K 得到最近樣本的依據，即預測新樣本屬於哪種 label 時找最近 K 個樣本 (鄰居)，再以多數決方法決定 label。此方法利用 sklearn package 的 neighbors.KNeighborsClassifier 進行實驗。

f) eXtreme Gradient Boosting, XGBoost: 為監督式學習方法，本質是 Gradient Boosting 方法，隨著建立不同棵樹來逐步降低殘差 (residual)，與隨機森林的差異點之一是，隨機森林的每棵樹用的資料不同 (從訓練資料重複抽樣固定比數)，而 XGBoost 用的資料相同。一個常見比較是: 隨機森林為深窄型、XGBoost 是淺寬型。此方法利用 xgboost package 的 XGBClassifier 進行實驗。

g) 降維技術: 常見如主成分分析 (Principal Component Analysis, PCA)，將原始特徵降低維度達到精簡資料的目的，本質上就是透過投影軸 (projection) 進行矩陣乘法。此方法利用 sklearn package 的 decomposition.PCA 進行實驗。

III. 實驗比較

針對每一個模型，先使用 pipeline 整合各部分，再使用 validation curve 選取合適的超參數 (hyperparameters)，並用 learning curve 檢驗是否有 high-bias 或 high-variance 問題。(所有 random state 均固定為 42)

A. 原始特徵與無降維

原始特徵主要使用 TABLE II 所列出 46 項特徵。

a) 邏輯斯迴歸 (Logistic Regression, LR): 主要調整懲罰係數 C。經過 validation curve，發現 C 的值無論大小均對結果沒有特別影響，故僅用預設值 (C = 1) 作為設定，如 Fig.2(a) 所示。利用 learning curve 可見 (見 Fig. 3(a))，整體來說隨著樣本數增加，邏輯斯迴歸模型大致呈現 low-bias 與 low-variance 的收斂現象。

b) 決策樹 (Decision Tree, DT): 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 6 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 6，如 Fig.2(b) 所示。利用 learning curve 可見 (見 Fig.3(b))，整體來說隨著樣本數增加，決策樹模型大致呈現 low-variance 的收斂現象。

TABLE II: 原始特徵

features	features	features	features
FlightDistance	IP_p0p2	ISO_SumBDT	$p1\_pt$
FlightDistanceError	IP_p1p2	$p0\_IsoBDT$	$p1\_p$
LifeTime	isolationa	$p1\_IsoBDT$	$p1\_eta$
IP	isolationb	$p2\_IsoBDT$	$p1\_IP$
IPSig	isolationc	$p0\_track\_Chi2Dof$	$p1\_IPSig$
VertexChi2	isolationd	$p1\_track\_Chi2Dof$	$p2\_pt$
dira	isolatione	$p2\_track\_Chi2Dof$	$p2\_p$
pt	isolationf	$p0\_pt$	$p2\_eta$
DOCAone	iso	$p0\_p$	$p2\_IP$
DOCBone	CDF1	$p0\_eta$	$p2\_IPsig$
DOCCone	CDF2	$p0\_IP$	
SPDhits	CDF3	$p0\_IPSig$	

[1]p0, p1, p2: Final States Tracks
[2]pt: Transverse Momentum
[3]p: Momentum
[4]eta: Pseudorapidity
[5]IP: Impact parameter
[6]IPSig: IP Significance
[7]DOCA: Distance of Closest Approach

c) 隨機森林 (Random Forest, RF): 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 6 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 6，如 Fig.2(c) 所示。利用 learning curve 可見 (見 Fig.3(c))，整體來說隨著樣本數增加，隨機森林模型大致呈現 low-variance 的收斂現象。(亦嘗試樹木數量 n_estimators 的 validation curve，但發現 50∼200 基本上毫無變動故省略)

d) AdaBoost: 主要調整 (弱) 分類器 n_estimators。經過 validation curve，發現 n_estimators 的值無論大小均對結果沒有特別影響，故僅用預設值 (n_estimators = 50) 作為設定，如 Fig.2(d) 所示。利用 learning curve 可見 (見 Fig.3(d))，整體來說隨著樣本數增加，AdaBoost 模型大致呈現 low-variance 的收斂現象。

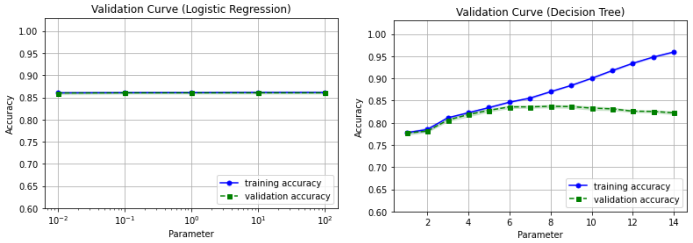
e) K-Nearest Neighbor, KNN: 主要調整最近鄰居數 n_neighbors。經過 validation curve，大體發現 n_neighbors = 6 後開始穩定，故僅取 n_neighbors = 6，如 Fig.2(e) 所示。利用 learning curve 可見 (見 Fig. 3(e))，整體來說隨著樣本數增加，K-Nearest 模型大致呈現 high-variance 的發散現象，顯然不適合做為最終模型。

f) eXtreme Gradient Boosting, XGBoost: 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 2 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 2，如 Fig.2(f) 所示。利用 learning curve 可見 (見 Fig.3(f))，整體來說隨著樣本數增加，XGBoost 模型大致呈現 low-variance 的收斂現象。(亦嘗試樹木數量 n_estimators 的 validation curve，但發現 50∼200 基本上毫無變動故省略)

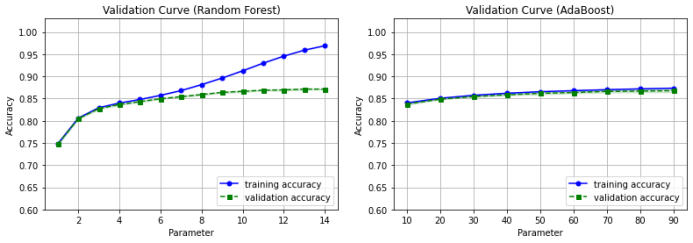
B. 原始特徵與有降維

根據 PCA 將訓練資料降維，在 Fig4 可以看出，針對累積變異 95% 作為門檻，其對應的壓縮特徵數為 28 個，故選取 features = 28 作為後續訓練。

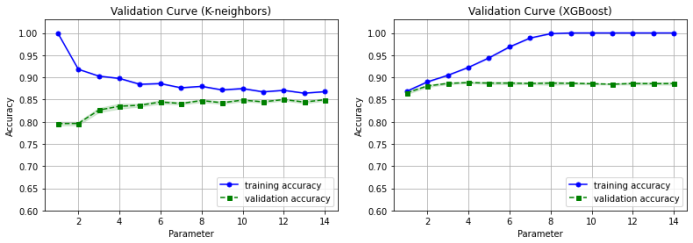
a) 邏輯斯迴歸 (Logistic Regression, LR): 主要調整懲罰係數 C。經過 validation curve，發現 C 的值無論大



(a) Logistic Regression (b) Decision Tree

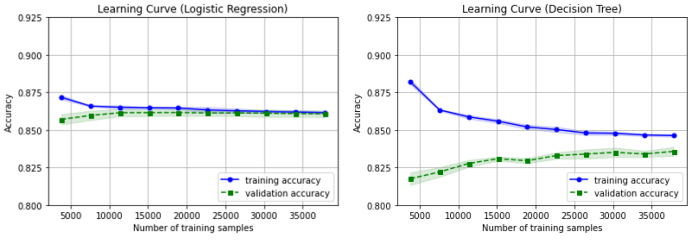


(c) Random Forest (d) AdaBoost

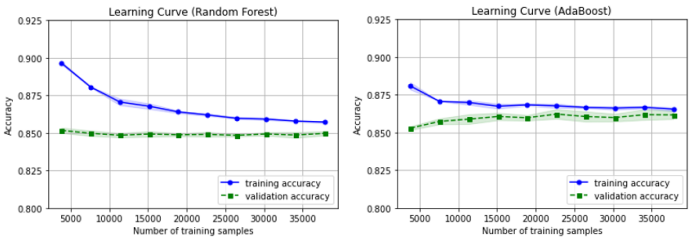


(e) K-neighbors (f) XGBoost

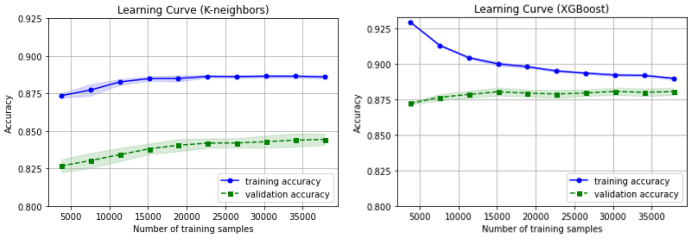
Fig. 2: 原始特徵與無降維之 validation curve



(a) Logistic Regression (b) Decision Tree



(c) Random Forest (d) AdaBoost



(e) K-neighbors (f) XGBoost

Fig. 3: 原始特徵與無降維之 learning curve

小均對結果沒有特別影響，故僅用預設值 (C = 1) 作為設定，如 Fig.5(a) 所示。利用 learning curve 可見 (見 Fig. 6(a))，整體來說隨著樣本數增加，邏輯斯迴歸模型大致呈現 low-bias 與 low-variance 的收斂現象。

b) 決策樹 (Decision Tree, DT): 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 6 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 6，如 Fig.5(b) 所示。利用 learning curve 可見 (見 Fig.6(b))，整體來說隨著樣本數增加，決策樹模型大致呈現 low-variance 的收斂現象。

c) 隨機森林 (Random Forest, RF): 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 6 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 6，如 Fig.5(c) 所示。利用 learning curve 可見 (見 Fig.6(c))，整體來說隨著樣本數增加，隨機森林模型大致呈現 low-variance 的收斂現象。(亦嘗試樹木數量 n_estimators 的 validation curve，但發現 50∼200 基本上毫無變動故省略)

d) AdaBoost: 主要調整 (弱) 分類器 n_estimators。經過 validation curve，發現 n_estimators 的值無論大小均對結果沒有特別影響，故僅用預設值 (n_estimators = 50) 作為設定，如 Fig.5(d) 所示。利用 learning curve 可見 (見 Fig.6(d))，整體來說隨著樣本數增加，AdaBoost 模型大致呈現 low-variance 的收斂現象。

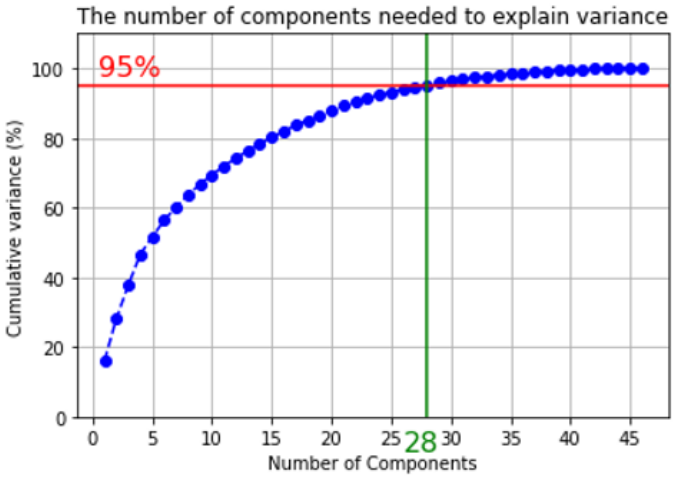
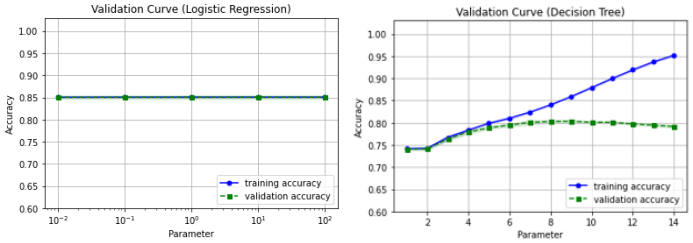
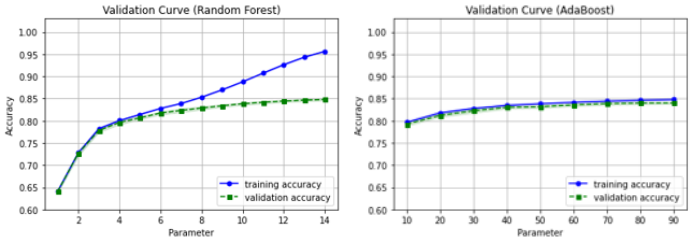


Fig. 4: 原始特徵使用 PCA 之累積變異圖

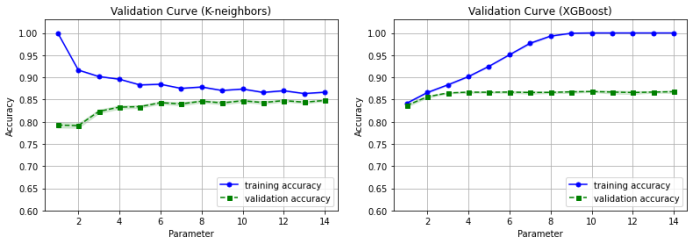
e) K-Nearest Neighbor, KNN: 主要調整最近鄰居數 n_neighbors。經過 validation curve，大體發現 n_neighbors = 6 後開始穩定，故僅取 n_neighbors = 6，如 Fig.5(e) 所示。利用 learning curve 可見 (見 Fig. 6(e))，整體來說隨著樣本數增加，K-Nearest 模型大致呈現 high-variance 的發散現象，顯然不適合做為最終模型。



(a) Logistic Regression (PCA) (b) Decision Tre (PCA)e

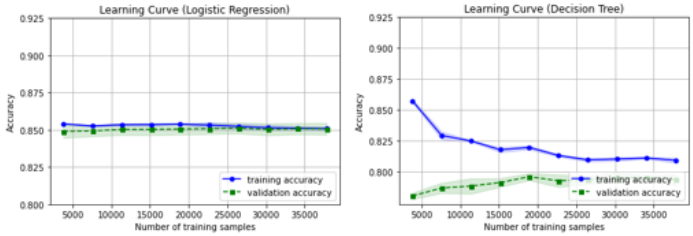


(c) Random Forest (PCA) (d) AdaBoost (PCA)

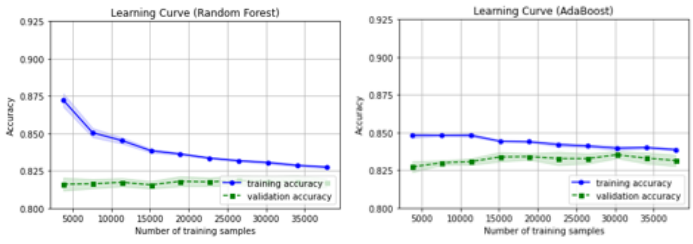


(e) K-neighbors (PCA) (f) XGBoost (PCA)

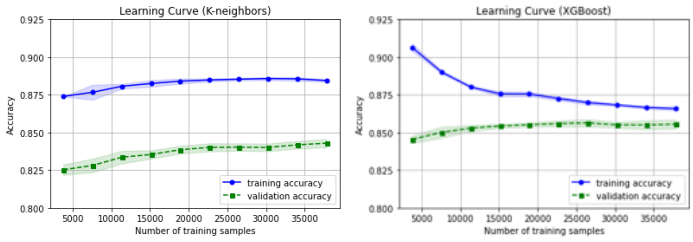
Fig. 5: 原始特徵與有降維之 validation curve



(a) Logistic Regression (PCA) (b) Decision Tree (PCA)



(c) Random Forest (PCA) (d) AdaBoost (PCA)



(e) K-neighbors (PCA) (f) XGBoost (PCA)

Fig. 6: 原始特徵與有降維之 learning curve

TABLE III: 原始特徵在訓練集之比較

Training	Acc.	Pre.	Sen.	Spe.	F1	AUC
LR	86.14	87.42	90.58	79.00	88.97	84.79
DT	84.56	87.81	87.05	80.55	87.43	83.80
RF	85.60	88.37	88.26	81.30	88.32	84.78
AdaBoost	84.78	88.30	90.17	80.77	89.23	85.47
KNN	88.65	89.97	91.84	83.52	90.90	87.68
XGBoost	88.71	89.96	91.96	83.47	90.95	87.72
LR(PCA)	85.08	86.43	89.93	77.27	88.15	83.60
DT(PCA)	81.12	83.83	85.98	73.30	84.89	79.64
RF(PCA)	82.64	82.74	90.80	90.80	90.80	80.14
AdaBoost(PCA)	83.63	85.22	88.88	75.18	87.01	82.03
KNN(PCA)	88.47	89.90	89.90	83.43	90.75	87.52
XGBoost(PCA)	86.47	87.55	91.00	79.16	89.24	85.08

TABLE IV: 原始特徵在測試集之比較

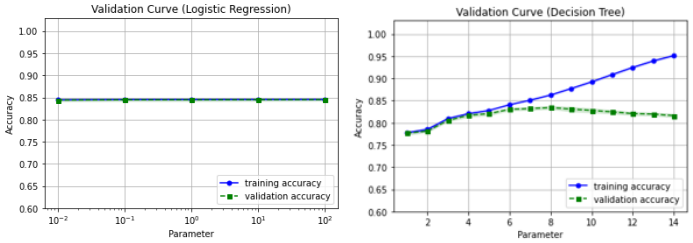
Test (%)	Acc.	F1	AUC	Agreement	Correlation
LR	85.85	88.74	84.47	✗	✓
DT	83.83	86.88	82.95	✗	✓
RF	84.91	87.77	84.03	✗	✓
AdaBoost	86.30	89.04	85.11	✗	✓
KNN	84.13	87.36	82.66	✗	✗
XGBoost	**87.97**	**90.36**	**86.91**	✗	✓
LR(PCA)	84.80	87.92	83.31	✗	✓
DT(PCA)	79.57	83.71	77.88	✓	✓
RF(PCA)	81.46	85.69	78.85	✓	✓
AdaBoost(PCA)	83.22	86.71	81.53	✗	✓
KNN(PCA)	84.08	87.30	82.65	✗	✗
XGBoost(PCA)	85.38	88.40	83.89	✗	✓

f) eXtreme Gradient Boosting, XGBoost: 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 2 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 2，如 Fig.5(f) 所示。利用 learning curve 可見 (見 Fig.6(f))，整體來說隨著樣本數增加，XGBoost 模型大致呈現 low-variance 的收斂現象。(亦嘗試樹木數量 n_estimators 的 validation curve，但發現 50∼200 基本上毫無變動故省略)

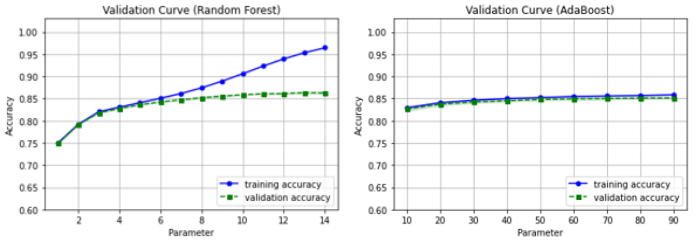
C. 新增特徵與無降維

新增特徵除了使用 TABLE II 所列出 46 項特徵外，還利用物理公式與 4 動量守恆定律 (Conservation of Four Momentum) 增加特徵，其增加特徵在 TABLE IV 列出。

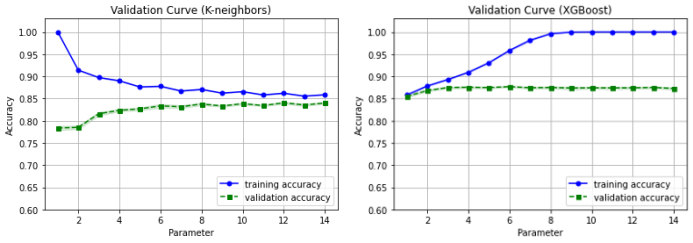
a) 邏輯斯迴歸 (Logistic Regression, LR): 主要調整懲罰係數 C。經過 validation curve，發現 C 的值無論大小均對結果沒有特別影響，故僅用預設值 (C = 1) 作為設定，如 Fig.7(a) 所示。利用 learning curve 可見 (見 Fig.



(a) Logistic Regression (b) Decision Tree

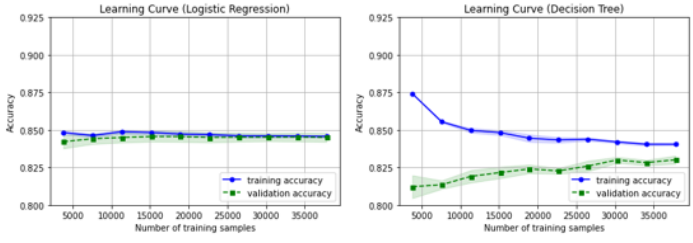


(c) Random Forest (d) AdaBoost

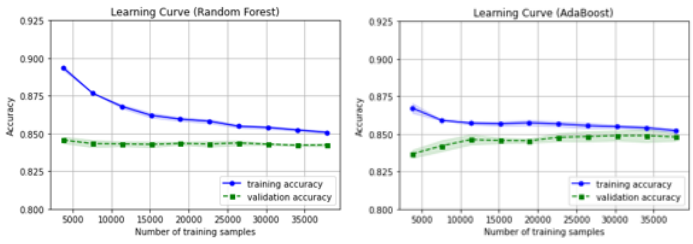


(e) K-neighbors (f) XGBoost

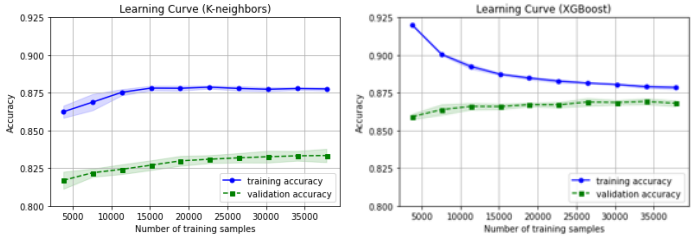
Fig. 7: 新增特徵與無降維之 validation curve



(a) Logistic Regression (b) Decision Tree



(c) Random Forest (d) AdaBoost



(e) K-neighbors (f) XGBoost

Fig. 8: 新增特徵與無降維之 learning curve

TABLE V: 新增特徵

new features	meaning	formula
$p0\_pz$ $p1\_pz$ $p2\_pz$ pz	momentum component along the z-axis	1. $p?_z = \sqrt{p^2 - p_t^2}$ 2. $p_z = p0_z + p1_z + p2_z$ (Conservation of four-momentum)
p	momentum	$p = \sqrt{p_t^2 - p_z^2}$
$measure\_speed$	speed at relativistic effects	$\frac{FlightDistance}{LifeTime}$

8(a))，整體來說隨著樣本數增加，邏輯斯迴歸模型大致呈現 low-bias 與 low-variance 的收斂現象。

b) 決策樹 (Decision Tree, DT): 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 6 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 6，如 Fig.7(b) 所示。利用 learning curve 可見 (見 Fig.8(b))，整體來說隨著樣本數增加，決策樹模型大致呈現 low-variance 的收斂現象。

c) 隨機森林 (Random Forest, RF): 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 6 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 6，如 Fig.7(c) 所示。利用 learning curve 可見 (見 Fig.8(c))，整體來說隨著樣本數增加，隨機森林模型大致呈現 low-variance 的收斂現象。(亦嘗試樹木數量 n_estimators 的 validation curve，但發現 50~200 基本上毫無變動故省略)

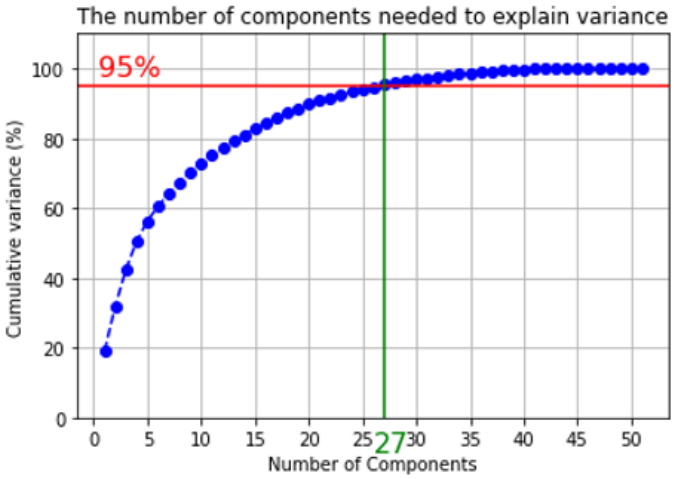
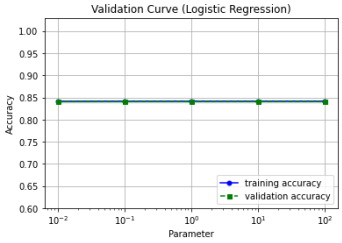
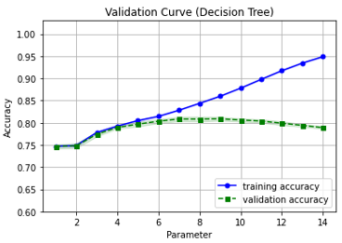


Fig. 9: 原始特徵使用 PCA 之累積變異圖

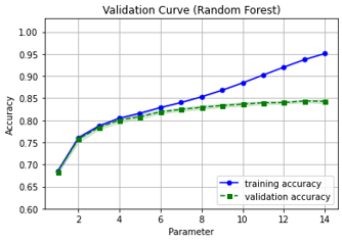
d) AdaBoost: 主要調整 (弱) 分類器 n_estimators。經過 validation curve，發現 n_estimators 的值無論大小均對結果沒有特別影響，故僅用預設值 (n_estimators = 50) 作為設定，如 Fig.7(d) 所示。利用 learning curve 可見 (見 Fig.8(d))，整體來說隨著樣本數增加，AdaBoost 模型大致呈現 low-variance 的收斂現象。



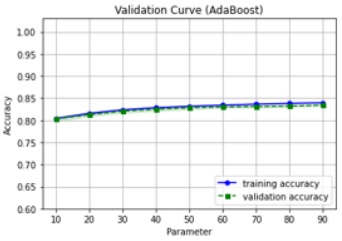
(a) Logistic Regression (PCA)



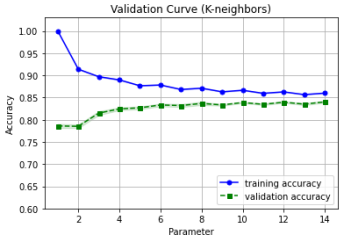
(b) Decision Tre (PCA)e



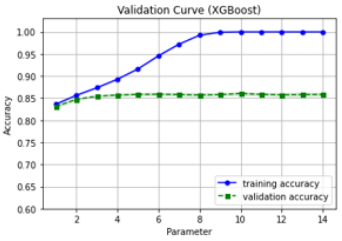
(c) Random Forest (PCA)



(d) AdaBoost (PCA)



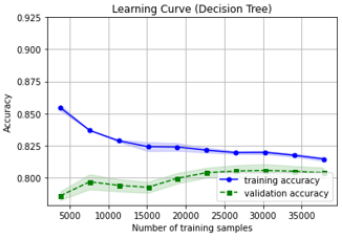
(e) K-neighbors (PCA)



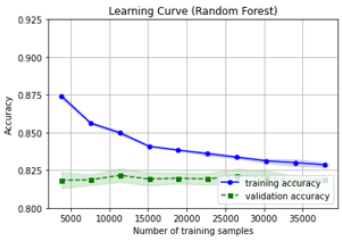
(f) XGBoost (PCA)

Fig. 10: 新增特徵與有降維之 validation curve

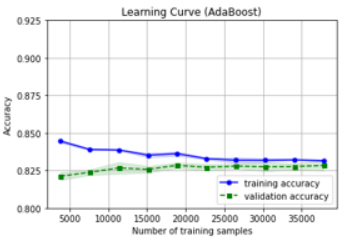
(a) Logistic Regression (PCA)



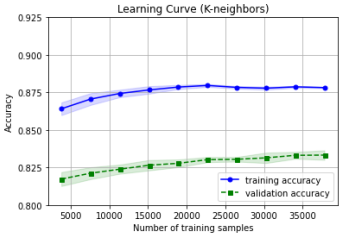
(b) Decision Tree (PCA)



(c) Random Forest (PCA)



(d) AdaBoost (PCA)



(e) K-neighbors (PCA)

(f) XGBoost (PCA)

Fig. 11: 新增特徵與有降維之 learning curve

TABLE VI: 新增特徵在訓練集之比較

Training (%)	Acc.	Pre.	Sen.	Spe.	F1	AUC
LR	84.56	85.90	89.71	76.28	87.76	82.99
DT	83.92	86.90	87.06	78.86	86.98	82.96
RF	84.96	88.10	87.43	80.99	87.76	84.21
AdaBoost	85.18	87.12	89.16	78.77	88.13	83.97
KNN	87.79	89.10	91.38	82.00	90.23	86.69
XGBoost	87.67	89.27	90.95	82.40	90.10	86.67
LR(PCA)	84.18	84.18	89.48	75.66	87.47	82.57
DT(PCA)	81.17	83.66	83.66	72.83	84.98	79.59
RF(PCA)	82.71	84.01	88.89	88.89	86.38	80.82
AdaBoost(PCA)	83.16	85.00	88.28	74.92	86.61	81.60
KNN(PCA)	87.89	85.58	91.35	82.31	90.30	86.83
XGBoost(PCA)	85.58	87.06	90.01	78.45	88.51	84.23

TABLE VII: 新增特徵在測試集之比較

Test (%)	Acc.	F1	AUC	Agreement	Correlation
LR	84.23	87.48	82.69	✓	✓
DT	82.97	86.22	81.92	✓	✓
RF	84.10	87.07	83.28	✓	✓
AdaBoost	84.52	87.62	83.22	✓	✓
KNN	83.53	86.89	82.03	✓	✗
XGBoost	**86.50**	**89.20**	**85.34**	✓	✓
LR(PCA)	84.00	87.33	82.36	✓	✓
DT(PCA)	79.70	83.87	77.93	✓	✓
RF(PCA)	81.57	85.58	79.41	✓	✓
AdaBoost(PCA)	82.76	86.33	81.09	✓	✓
KNN(PCA)	83.35	86.73	81.85	✓	✗
XGBoost(PCA)	84.64	87.78	83.20	✓	✓

e) K-Nearest Neighbor, KNN: 主要調整最近鄰居數 n_neighbors。經過 validation curve，大體發現 n_neighbors = 6 後開始穩定，故僅取 n_neighbors = 6，如 Fig.7(e) 所示。利用 learning curve 可見 (見 Fig. 8(e))，整體來說隨著樣本數增加，K-Nearest 模型大致呈現 high-variance 的發散現象，顯然不適合做為最終模型。

f) eXtreme Gradient Boosting, XGBoost: 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 2 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 2，如 Fig.7(f) 所示。利用 learning curve 可見 (見 Fig.8(f))，整體來說隨著樣本數增加，XGBoost 模型大致呈現 low-variance 的收斂現象。(亦嘗試樹木數量 n_estimators 的 validation curve，但發現 50~200 基本上毫無變動故省略)

D. 新增特徵與有降維

根據 PCA 將訓練資料降維，在 Fig9 可以看出，針對累積變異 95% 作為門檻，其對應的壓縮特徵數為 27 個，故選取 features = 27 作為後續訓練。

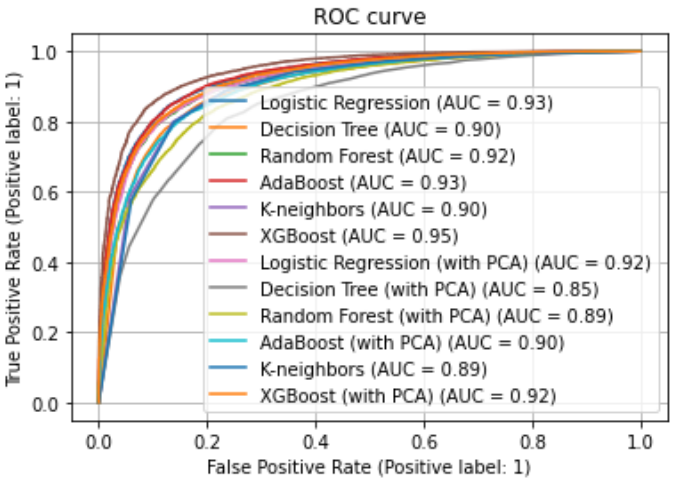


Fig. 12: 原始特徵之各種機器學習模型 AUC 比較圖

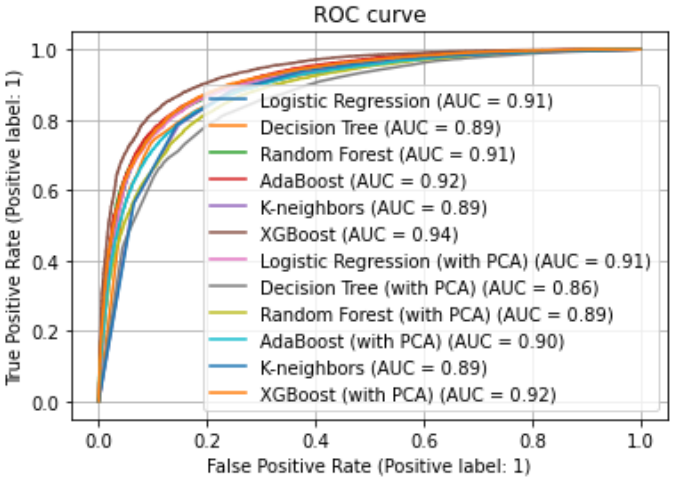


Fig. 13: 新增特徵之各種機器學習模型 AUC 比較圖

a) 邏輯斯迴歸 *(Logistic Regression, LR):* 主要調整懲罰係數 C。經過 validation curve，發現 C 的值無論大小均對結果沒有特別影響，故僅用預設值 (C = 1) 作為設定，如 Fig.10(a) 所示。利用 learning curve 可見 (見 Fig. 11(a))，整體來說隨著樣本數增加，邏輯斯迴歸模型大致呈現 low-bias 與 low-variance 的收斂現象。

b) 決策樹 *(Decision Tree, DT):* 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 6 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 6，如 Fig.10(b) 所示。利用 learning curve 可見 (見 Fig.11(b))，整體來說隨著樣本數增加，決策樹模型大致呈現 low-variance 的收斂現象。

c) 隨機森林 *(Random Forest, RF):* 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 6 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 6，如 Fig.10(c) 所示。利用 learning curve 可見 (見 Fig.11(c))，整體來說隨著樣本數增加，隨機森林模型大致呈現 low-variance 的收斂現象。(亦嘗試樹木數量 n_estimators 的 validation curve，但發現 50~200 基本上毫無變動故省略)

TABLE VIII: SVM 結合 K-means 在測試集之比較

Test	Acc.	Pre.	Sen.	Spe.	F1
K = 3 (原始特徵)	74.53	76.40	84.96	57.73	80.45
K = 4 (原始特徵)	73.86	74.85	86.78	53.05	80.37
K = 5 (原始特徵)	70.56	69.07	94.69	31.71	79.87
K = 3 (新增特徵)	74.60	77.28	83.32	60.55	80.19
K = 4 (新增特徵)	73.91	74.80	87.03	52.79	80.45
K = 5 (新增特徵)	67.19	76.30	67.92	66.02	71.86

TABLE IX: 訓練模型之時間比較

Time (s)	LR	DT	RF	AdaBoost	KNN	XGBoost
原始特徵						
無降維	0.660	1.180	9.308	10.695	0.078	2.562
PCA 降維	0.233	0.959	9.608	8.987	0.195	2.147
新增特徵						
無降維	1.010	1.397	11.264	12.365	0.063	2.136
PCA 降維	0.257	0.985	9.665	8.067	0.145	1.589

基本上毫無變動故省略)

d) AdaBoost: 主要調整 (弱) 分類器 n_estimators。經過 validation curve，發現 n_estimators 的值無論大小均對結果沒有特別影響，故僅用預設值 (n_estimators = 50) 作為設定，如 Fig.10(d) 所示。利用 learning curve 可見 (見 Fig.11(d))，整體來說隨著樣本數增加，AdaBoost 模型大致呈現 low-variance 的收斂現象。

e) K-Nearest Neighbor, KNN: 主要調整最近鄰居數 n_neighbors。經過 validation curve，大體發現 n_neighbors = 6 後開始穩定，故僅取 n_neighbors = 6，如 Fig.10(e) 所示。利用 learning curve 可見 (見 Fig. 11(e))，整體來說隨著樣本數增加，K-Nearest 模型大致呈現 high-variance 的發散現象，顯然不適合做為最終模型。

f) eXtreme Gradient Boosting, XGBoost: 主要調整最大深度 max_depth。經過 validation curve，發現 max_depth = 2 後開始分開 (overfitting 越來越明顯) 因此設定 max_depth = 2，如 Fig.10(f) 所示。利用 learning curve 可見 (見 Fig.11(f))，整體來說隨著樣本數增加，XGBoost 模型大致呈現 low-variance 的收斂現象。(亦嘗試樹木數量 n_estimators 的 validation curve，但發現 50~200 基本上毫無變動故省略)

E. 實驗小結

根據上述實驗，利用原始特徵 (TABLE II) 在訓練集與測試集之效能列在 TABLE III、TABLE IV，XGBoost 總體最好但無法通過 Agreement 檢驗，故嘗試新增特徵 (除了 TABLE II 外，再加上 TABLE V，此外刪除'SPDhits' 特徵) 提升通過可能性；新增特徵後，在在訓練集與測試集之效能列在 TABLE VI、TABLE VII，明顯看出效能提升且同時通過 Agreement、Correlation 檢驗，因此新增特徵是有意義的。(物理公式由主辦方說明文件提供，此外新增特徵之' 測量速度' 是在相對論效應/Relativistic Effects 下的結果)

F. K-means 與 SVM

本小節主要討論如何有效率使用 SVM 進行預測，其使用訓練資料集上述表現最好之「新增特徵且不降維」。如何訓練並驗證 SVM，其步驟如下:

Step 1. 使用 K-means 找出 centroids (elbow method 決定群數 K)

Step 2. 找出訓練集離 centroids 最接近的樣本作為新訓練集 (size = K)

Step 3. 使用 SVM 模型訓練，並作為預測樣本的結果 (類似 query)

Step 4. 測試集透過 Step 1 找出歸屬類別，即能知道其 label 為何

如 TABLE VIII 所示，K = 3 表現較好 (亦可用 elbow method 驗證)，但也不能同時通過 Agreement、Correlation 檢驗。

IV. 討論與結論

LVF 的發現是有力懷疑標準模型的證據之一，迄今為止，無論是 $\tau^- \to \mu^- + \mu^- + \mu^+$ $D^+_S \to \pi^- + \phi$ ($\to \mu^- + \mu^+$) 衰變僅只是可能存在 (其衰變過程機率非常低)，仍需大量實驗來確認。本次競賽透過機器學習方法幫助實驗物理學家將數據進行有效過濾以專注於 $\tau^- \to \mu^- + \mu^- + \mu^+$ 的可能數據，此外，根據專家建議必須分類器同時考慮真實數據與模擬數據的分布應該相似 (Agreement-test) 與分類依據應該與質量無關 (Correlation-test)。實驗得知新增特徵 (物理學相關特徵) 使用 XGBoost 方法 (不降維) 表現最好，其 accuracy 約 86.5%、F1 score 約 89.2%、AUC 約 85.34，同時通過題目所要求的 Agreement、Correlation 檢驗，符合題目要求。Fig. 12、Fig. 13 分別將原始特徵 (46 項特徵)、新增特徵 (45 + 6 項特徵，經實驗發現'SPDhits' 特徵加入反而無法通過 Agreement、Correlation 檢驗) 畫出 AUC，可以看出新增特徵、XGBoost 與不降維表現最佳，同時在 TABLE IX 顯示訓練模型的所需時間比較，降維 (使用 PCA) 略為時間下降，但並非所有，而 AdaBoost 耗時最久、KNN 耗時最短。截至目前，$\tau^- \to \mu^- + \mu^- + \mu^+$ 的衰變過程依舊持續驗證 [7](2022, CERN)，標準模型的初步完善不代表就此找出萬能解釋 (標準模型仍舊不能對萬有引力/gravitation 進行解釋，所謂自旋數 = 2 的重力子/graviton 在實驗上沒有找到)，恰好相反，標準模型的問題日益增多，這也是現在物理學家們想要改善的部分之一 – 找出超越標準模型的證據，對標準模型進行修正甚至改寫，許多學者願意花費一生找到能完美解釋宇宙萬物的理論且能與實驗契合，同時粒子物理的研究孕育出許多諾貝爾物理學獎得主，也許不久的未來真的能找出萬有理論 (Theory of Everything, ToE) 值得期待。(題外話，2022 年的 natrue 刊登的論文顯示，蒐集並分析 10 年數據，W boson 的質量竟然與理論預測不符合![9])

Acknowledgment

感謝彭老師這學期的「機器學習」教導，從基礎理論 (例如 APEC、No free lunch theorem 等) 到應用介紹 (例如 logistic regression、Adaboot、PCA 等)。本次期末專題做麻煩的是排版 (初次使用 Latex)，光是排版要整齊就花費不少時間，比 Word 輕鬆在不用編號，但是 [htbp] 排列組合很久，機器學習方法頗像煉丹，因為要調整 hyperparameters，利用 grid search、random search 是其中一種方法，但是遇到資料量不但跑很久並且會讓電腦當機...，故本次實驗僅針對單一 hyperparameter，另一方面其他參數較無變化所以只挑變化較大、較常見的 hyperparameter 為主。

References

[1] https://www.kaggle.com/competitions/flavours-of-physics
[2] Kelvin, Lord. (1901). I. Nineteenth century clouds over the dynamical theory of heat and light. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(7), 1–40. doi:10.1080/14786440109462664
[3] https://zh.wikipedia.org/zh-tw
[4] https://pdglive.lbl.gov/Viewer.action
[5] https://scikit-learn.org
[6] https://xgboost.readthedocs.io
[7] https://cds.cern.ch/record/2806235
[8] https:// www.amazon.co.uk/ Standard-Lagrangian-Particle-Physics-T-Shirt/dp/B07TJ1LPQH
[9] Davide Castelvecchi and Elizabeth Gibney. (2022). Particle' s surprise mass threatens to upend the standard model. natrue, 604(7905), 225-226. doi: 10.1038/d41586-022-01014-5
[10] http:// www.thomasgmccarthy.com/ an-introduction-to-collider-physics-ix
[11] https://indico.cern.ch/event/92209/contributions/2114409/attachments/1098701/1567290/CST2010-MC.pdf

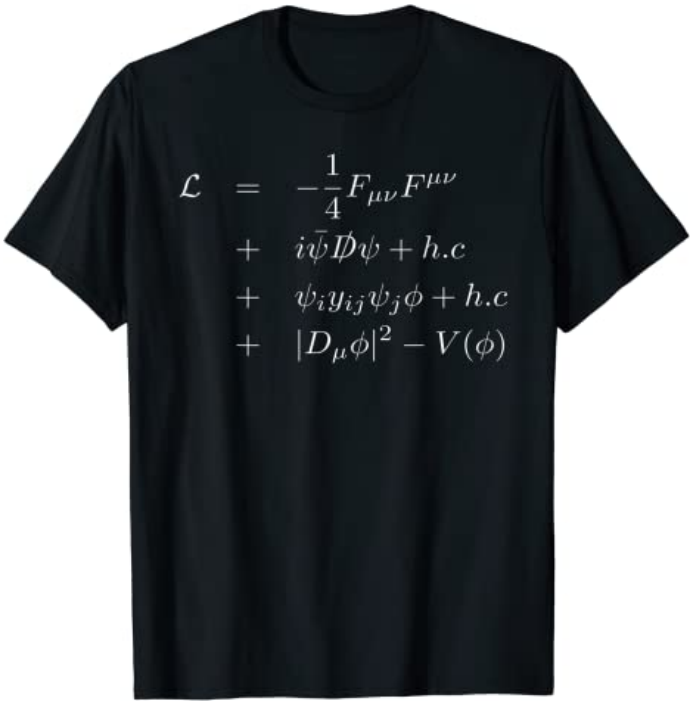


Fig. 14: Standard Model Lagrangian of Particle Physics