

An artificial intelligence model for heart disease detection using machine learning algorithms

Victor Changa, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, MA Hossain

Healthcare Analytics Volume 2, November 2022

Group 6 Member:

313707002 羅芷羚
313707013 簡夢萱

313707011 涂鉉挺
313707014 陳紀蓁



Agenda

- ▶ 論文介紹
- ▶ 資料庫與前處理
- ▶ 模型介紹
- ▶ 模型結果
- ▶ 特徵重要性
- ▶ 結論



論文介紹



主題新穎性



- 以 Python + Machine Learning 建立心臟疾病偵測模型
 - 將 ML 模型導入臨床決策輔助情境，符合近年智慧醫療研究趨勢
 - 醫療資料分析
 - 臨床決策輔助系統（CDSS）
- 將 AI 模型定位為「預警系統」
 - 聚焦 CAD（冠狀動脈疾病）早期無症狀階段
- 跨領域整合
 - 醫療臨床資料
 - 機器學習模型
 - 醫療資訊安全與隱私（HIPAA）





實際應用可行性

- 資料層面：資料可取得性高
 - 資料集的特徵包含年齡、性別、血壓、膽固醇等臨床常見指標
 - 不依賴高成本或高度侵入性檢查
- 模型層面：可重現性高
 - Random Forest - 不需大量超參數調整、對小樣本資料具穩定性、不易過度擬合 (overfitting)
- 系統層面：運算成本低
 - Python 為開源語言、可直接部署於醫療資訊單位或研究機構
 - 不需高效能運算資源，適合中小型醫療機構
- 資安與法規層面：合規性高
 - Python 支援 資料加密、網路安全、存取控制
 - 符合 HIPAA 醫療資料保護要求





研究方法創新性



- 研究方法架構 - 標準 Machine Learning pipeline:
 - Exploratory Data Analysis (EDA)
 - 資料前處理 - Data Cleaning, Encoding, Scaling
 - 建模 - KNN, Decision Tree, SVM, Logistic Regression, **Random Forest**
 - 模型評估 - 10-fold Cross-Validation, Accuracy, Precision, Recall, F1-score, Confusion Matrix
- 結構化比較多種模型結果
- 實證證明 Random Forest 在醫療資料上的穩定性與預測效能高



資料庫與前處理

資料庫

- 美國加州大學歐文分校機器學習資料庫
 - Heart Disease Cleveland 資料集
- 筆數：304筆
- 特徵變數：14個

特徵變數

英文變數	中文名稱	型態	說明
sex	性別	類別（二元）	1 = 男性，0 = 女性
cp	胸痛類型	類別（四類）	0–3：典型心絞痛、非典型心絞痛、非心絞痛性疼痛、無症狀
fbs	空腹血糖 >120 mg/dl	類別（二元）	1 = 是、0 = 否
restecg	心電圖結果	類別（三類）	0, 1, 2 三種 ECG 型態
exang	運動誘發性心絞痛	類別（二元）	1 = 有、0 = 無
slope	ST 斜率	類別（三類）	0, 1, 2 : ST segment 的斜率（上升/平坦/下降）
ca	主血管數量	類別（四類）	螢光顯影顯示之狹窄血管數（0–3 條）
thal	心肌灌流掃描結果	類別（三類）	正常、固定缺陷、可逆缺陷等
age	年齡	連續	受試者年齡（歲）
trestbps	血壓	連續	靜止時的收縮壓（mm Hg）
chol	血液中總膽固醇	連續	mg/dl
thalach	運動最大心率	連續	運動測試中達到的最大心率（bpm）
oldpeak	ST 段壓低	連續	運動相較於靜止時引起的 ST 壓低幅度（反映心肌缺氧）
target	心臟病（目標變數）	類別（二元）	1 = 有心臟病，0 = 無心臟病

資料前處理

- 刪除缺失值
 - 資料筆數：297
- 類別變數
 - One-Hot Encoding 轉換為二元向量
- 連續變數
 - StandardScaler 進行標準化
- 10-fold Cross-Validation 進行交叉驗證
 - Stratified random sampling

模型介紹

論文模型

KNN

- 模型優點
 - 概念直觀、無須進行複雜模型訓練
 - 對於非線性資料有良好表現
- 模型缺點
 - 特徵維度高時，運算成本偏高
 - 對資料尺度敏感

Decision Tree

- 模型優點
 - 模型具高可解釋性
 - 能處理非線性資料
- 模型缺點
 - 資料中的雜訊較為敏感
 - 單棵決策樹穩定性相對有限

Random Forest

- 模型優點
 - 準確率高
 - 引入隨機性，不易過擬合
 - 適用於分類與迴歸問題
- 模型缺點
 - 訓練時需要大量記憶空間與時間

Logistic Regression

- 模型優點
 - 計算速度快且簡單
 - 可解釋性強
- 模型缺點
 - 只能處理線性關係
 - 對異常值敏感

論文模型

SVM (Linear) SVM (RBF)

- `kernel = linear`
- `kernel = rbf`
- 模型優點
 - 適合處理高維度資料
 - 有良好的泛化能力與穩定性
- 模型缺點
 - 不適合樣本數非常大的資料集

我們新加入的模型

XGBoost

- 模型優點
 - 泛化能力佳、不易過度擬合
 - 非線性資料中具備良好預測能力
- 模型缺點
 - 模型結構複雜，可解釋性較低
 - 計算時間與記憶體資源需求較大

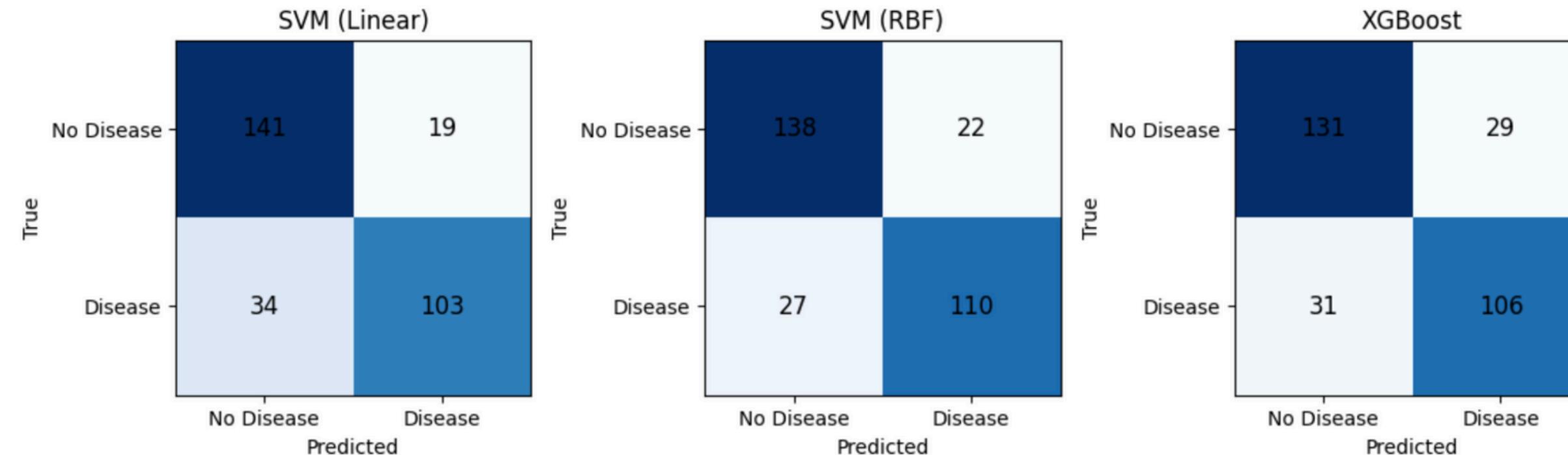
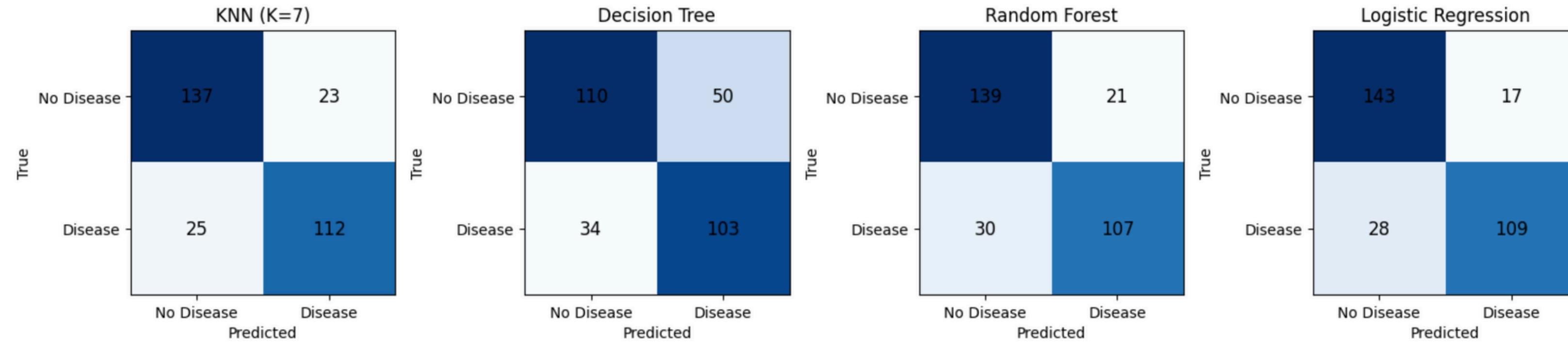
模型結果

模型結果

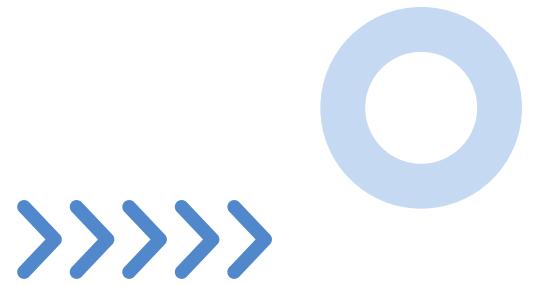
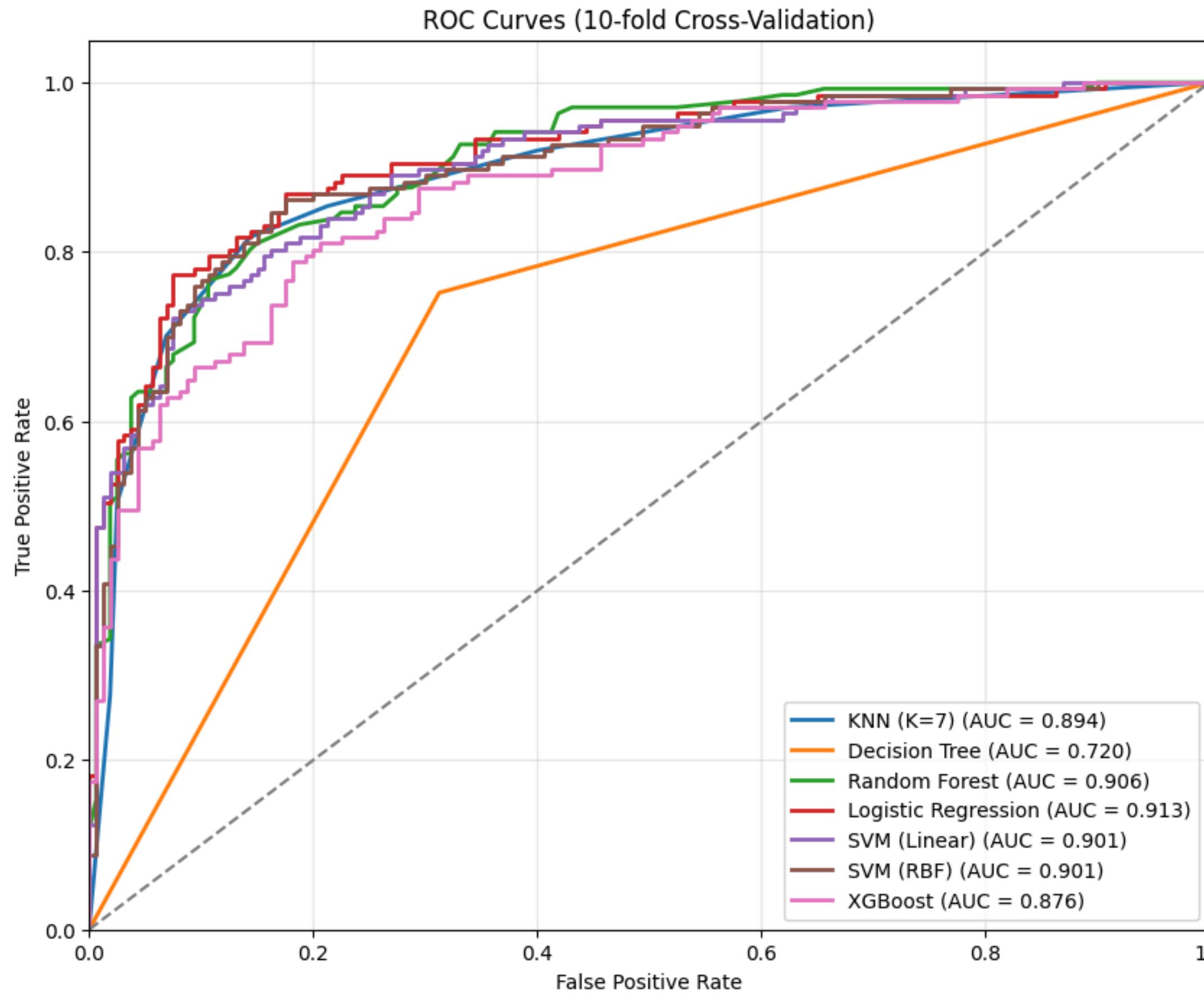
來源	Model	小組實證結果				論文實證結果
		Accuracy	Precision	Recall	F1	
論文模型	KNN (K=7)	84%	84%	82%	83%	87%
	Decision Tree	72%	67%	75%	71%	79%
	Random Forest	83%	84%	78%	81%	84%
	Logistic Regression	85%	87%	80%	83%	NA
	SVM (Linear)	82%	86%	75%	80%	83%
	SVM (RBF)	84%	84%	80%	82%	
新增模型	XGBoost	80%	80%	78%	78%	NA

各方法小組實證與論文結果類似

Confusion Matrices



ROC Curves



Grid Search 方法



>>>

隨機森林	
預設	Grid Search
max_depth = None	"model__max_depth": [None, 5, 10]
min_samples_leaf = 1	"model__min_samples_leaf": [1, 5, 10]

XGBoost	
預設	Grid Search
n_estimators=200	"model__n_estimators": [200, 400, 800]
max_depth=3	"model__max_depth": [2, 3, 4]
learning_rate=0.1	"model__learning_rate": [0.01, 0.05, 0.1]

Grid Search 結果



Model	Accuracy	Precision	Recall	F1
Random Forest (Baseline)	83%	84%	78%	81%
Random Forest (Tuned)	84%	86%	79%	82%
XGBoost (Baseline)	80%	80%	78%	78%
XGBoost (Tuned)	85%	86%	80%	83%

新增的模型（XGBoost）效能將優於論文所主張之核心（Random Forest）方法好

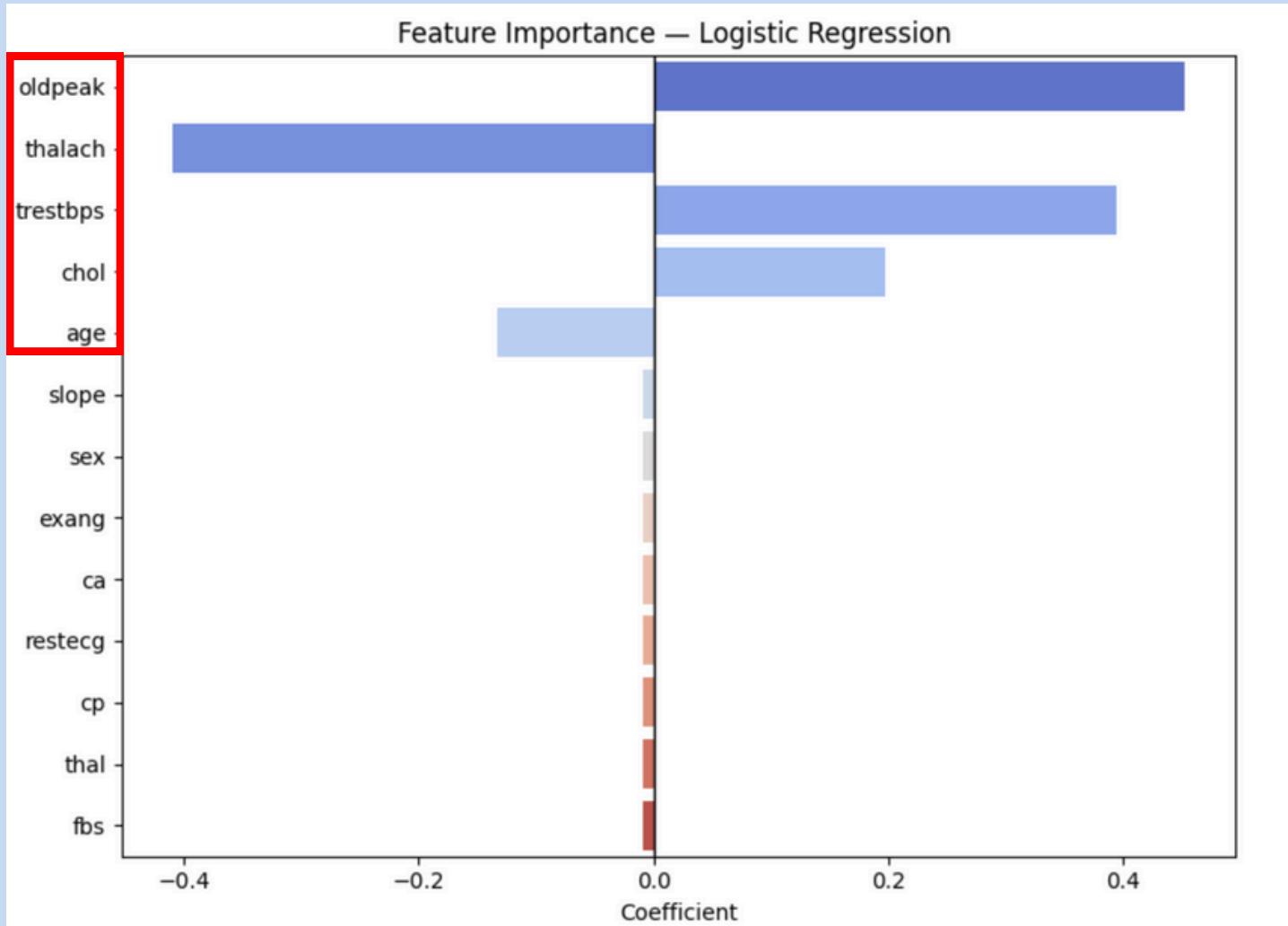
Feature importance & SHAP

衡量不同特徵對模型預測結果影響程度的指標

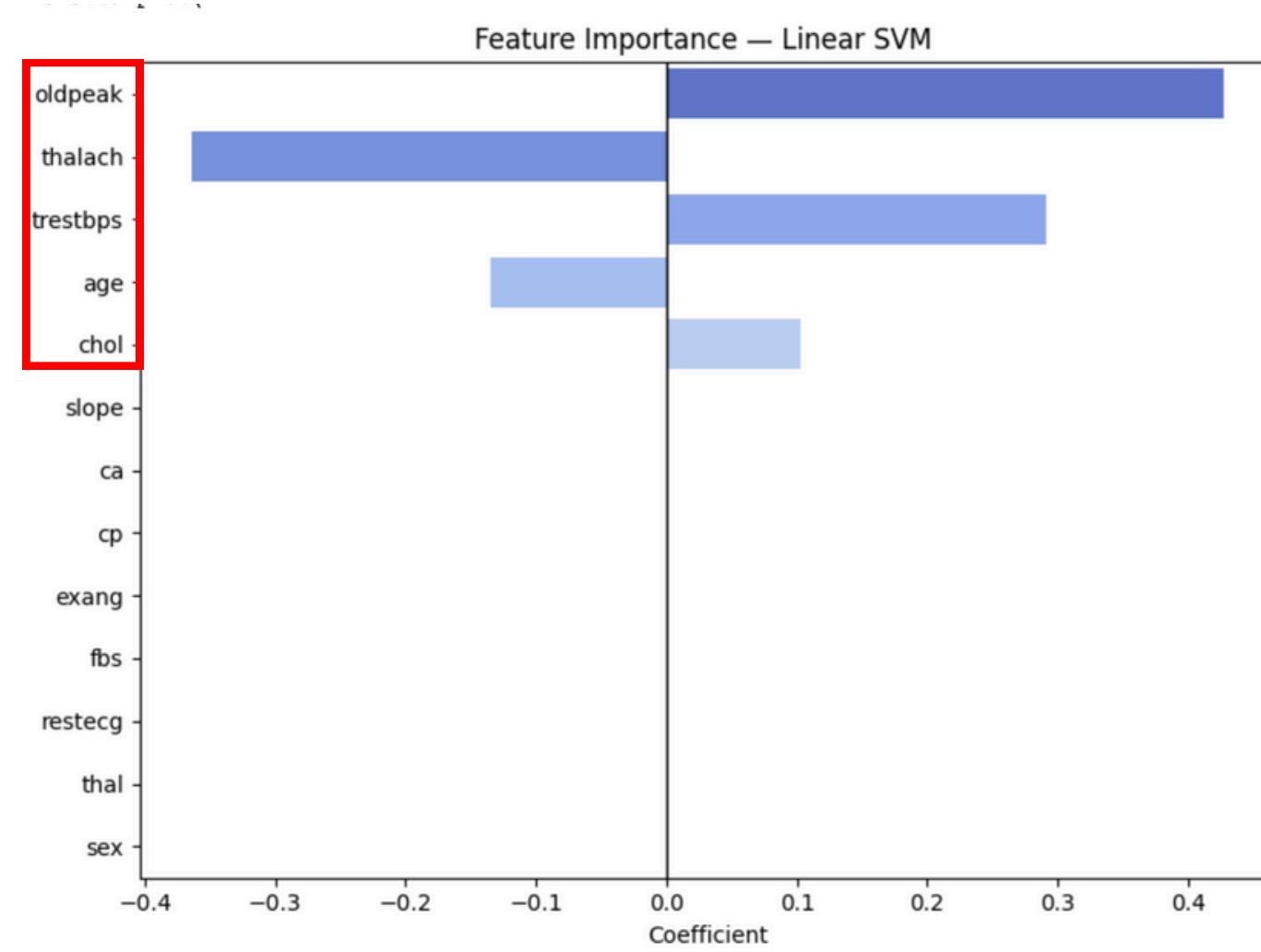


線性模型

I . Logistic Regression



II . SVM(Linear)



正相關:

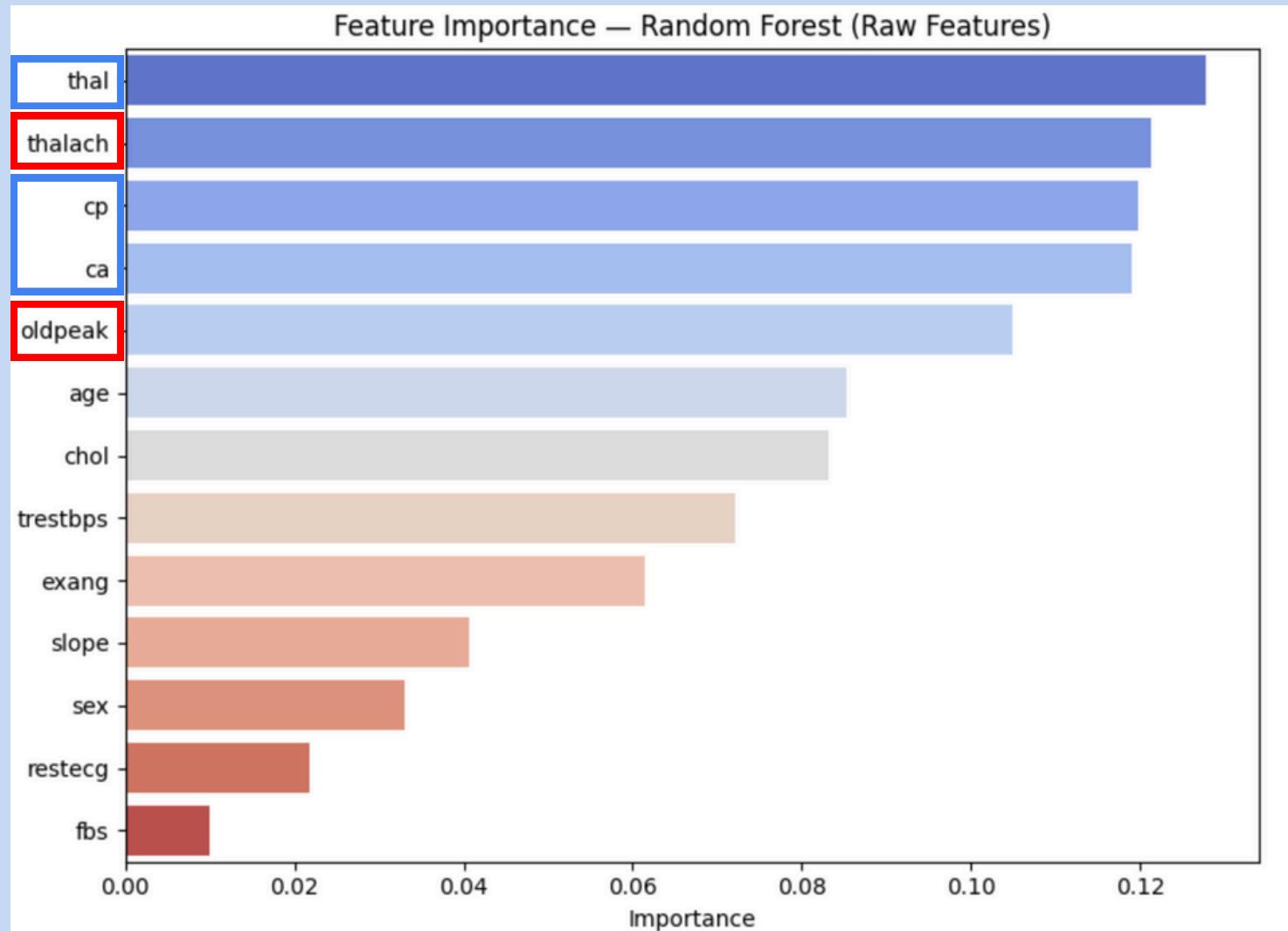
- oldpeak: 與休息狀態相比，運動導致了心電圖中的 ST 段出現下移（心臟在壓力（運動）下出現心肌缺血）
- trestbps: 身體完全放鬆、靜止狀態下測量到的血壓
- chol: 血液中的總膽固醇

負相關:

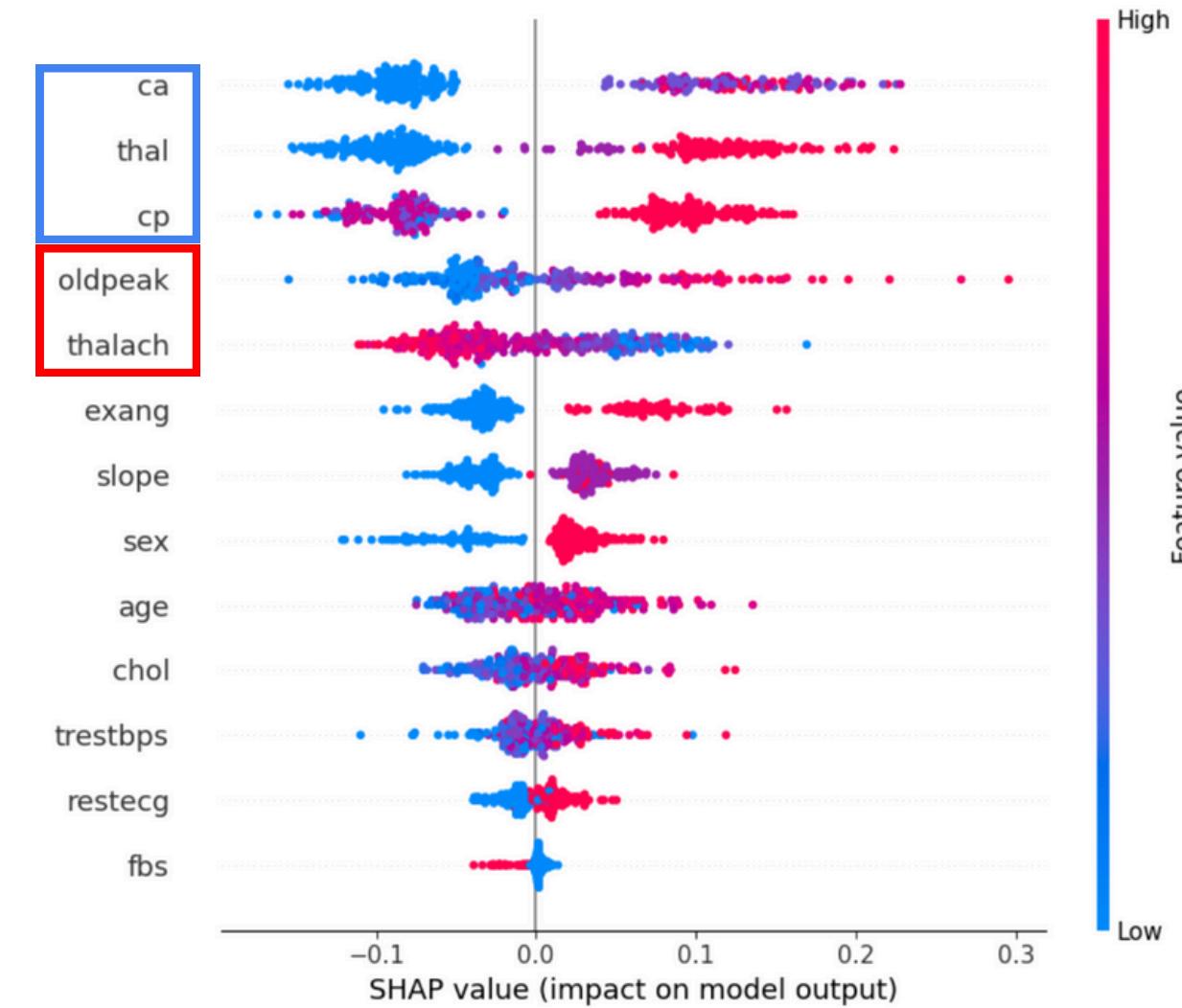
- thalach: 運動中所達到的最高心跳率
- age: 年齡

III. Random Forest

Feature Importance



SHAP



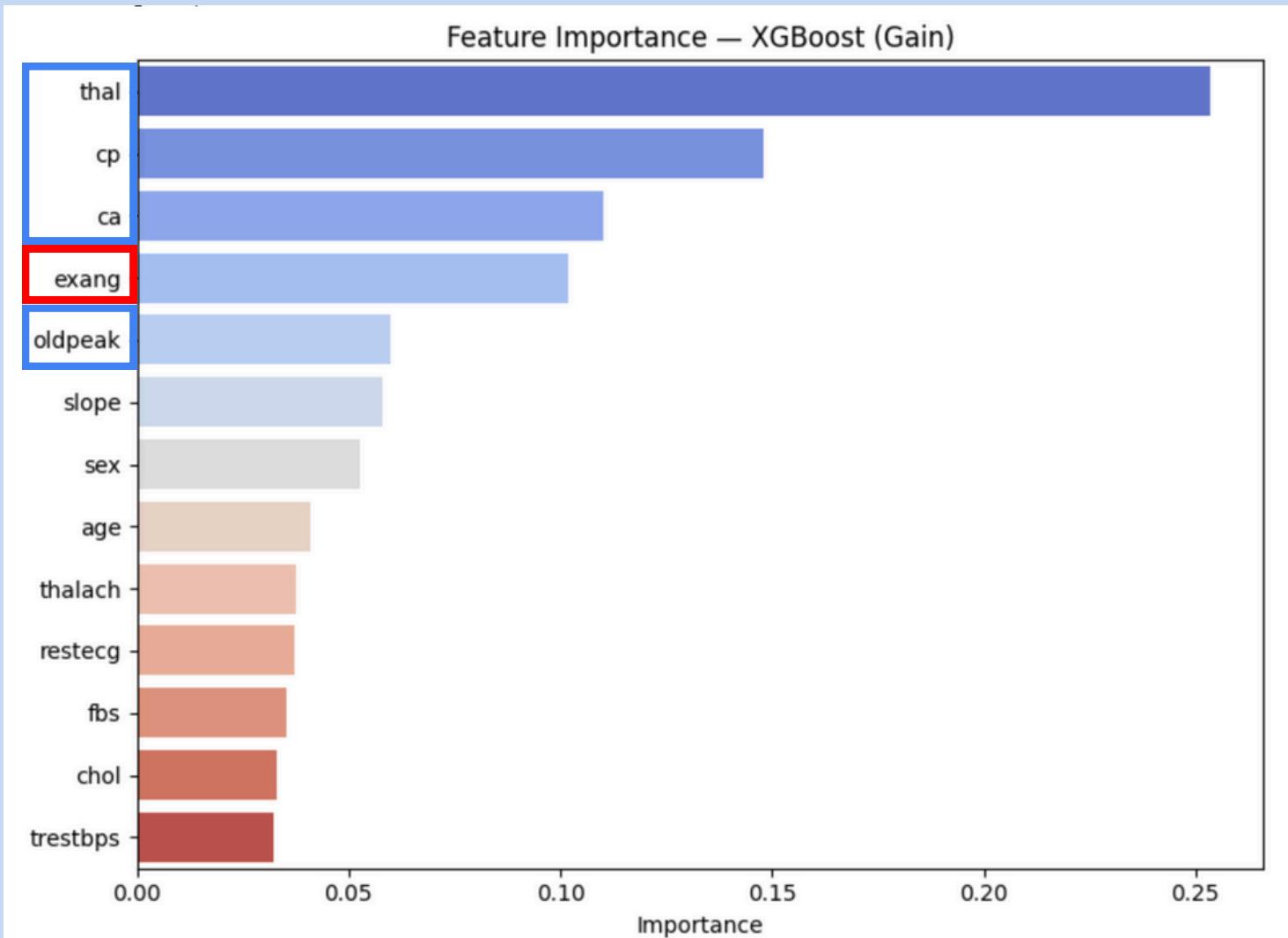
正相關：

- thal: 血液流向心臟的情況
- cp: 胸痛類型
- ca: 透視檢查顯色的主要血管數量，顯色異常通表血管狹窄或有堆積

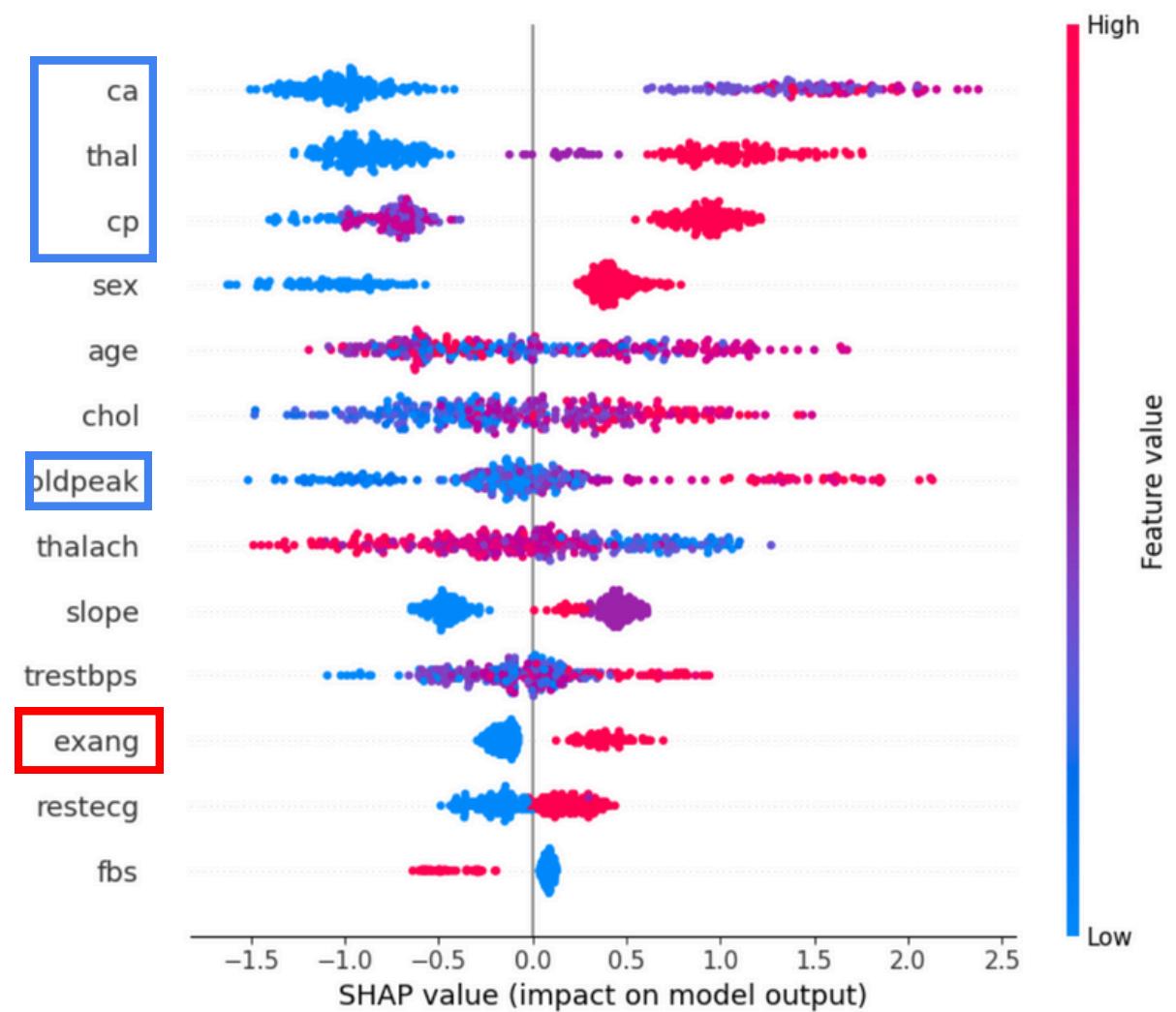


IV. XGBoost

Feature Importance



SHAP



正相關：

- exang：運動誘發心絞痛





健檢複查指標

- **一致重要指標 (每半年複檢一次)**
oldpeak、thalach
- **線性模型關鍵指標 (每一年複檢一次)**
trestbps、chol、age
- **樹狀模型關鍵指標 (每兩年複檢一次)**
thal、cp、ca



Conclusion

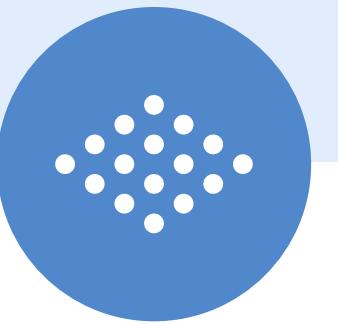
結論



結論

- 完整重現論文中所採用的模型架構與實驗流程，包括 KNN、SVM、Logistic Regression、Decision Tree 以及 Random Forest
- 三項延伸
 1. 加入論文中未使用的 XGBoost 模型，比較和原始模型的表現
 2. 透過 GridSearch 進行參數調整，以提升模型效能
 3. 結合 Feature Importance 與 SHAP 分析，強化模型的可解釋性
- 進一步將模型結果轉化為具體可行的健檢複查指標，讓研究成果能夠從模型比較，延伸至實際健康管理與風險評估的應用情境





**THANK
YOU!**

