* 模型

支援向量機 (SVM)

是一種監督式學習，透過將資料點映在高維空間中並建構超平面，以進行分類、迴歸或其他任務。給定一組訓練集，每個資料點會被標記為屬於兩個類別中的其中一類，支援向量機假設存在一個超平面

可以完美分割兩組資料，透過計算 w 及 b 使得兩類之間的邊界最大化，並將資料重新分類為兩個類別之一。值得注意的是，需要先對資料的每個特徵進行標準化，使資料點的每個特徵在支援向量機裡同等重要。

隨機森林 (Random forest )

屬於集成學習 (Ensemble learning)法。訓練分類器時，建造多個分類樹(Decision Tree)，再以拔靴法 (Bootstrap) 抽樣進行訓練，最後採多數決決定分類結果。樹的分支則是利用熵 (Entropy)或吉尼不純度 (Gini impurity)，尋找特徵子集內擁有最大資訊增益(Information Gain, IG)的變數，進而停止分割節點。與單一分類樹相比，此方法較能降低偏誤及變異，且不易過度擬和。

Rusboost

全名為Random Under Sampling ，用於解決不平衡資料的分類器，RUSboost 結合資料抽樣和集成模型的方法。其步驟為：先給定所有變數的初始權重，透過不斷遞迴計算，更新每個變數的權重，其中每次計算會隨機抽取一定數量的多數類樣本，和所有的少數類組成訓練數據集，接著利用決策樹分類，計算出錯誤率，針對錯誤最多的類別改變權重計算方式，在固定迴圈或收斂下，決定目前為止最佳的分類器作為分類結果。

單純貝氏(Naive bayes)

為一種構建分類器的方法。它不是訓練這種分類器的單一演算法，而是一系列基於相同原理的演算法：所有單純貝氏分類器都假定樣本特徵間彼此獨立。在許多實際應用中，單純貝氏模型參數估計使用最大概似估計方法，換言之，在不用到貝氏機率或者任何貝氏模型的情況下，單純貝氏模型也能奏效。儘管假設很簡單，但單純貝氏分類器在很多複雜的現實情形中仍能夠取得相當好的效果。優勢在於只需根據少量的訓練資料，即能估計出必要的參數（變數的均值和變異數），且由於變數獨立假設，所以不需要確定整個共變異數矩陣。

K-nearest neighbor

又稱KNN演算法，屬監督式學習。訓練資料為具有標籤的資料，計算每個樣本點與目標點之歐式距離，選k個最近點並做類別判斷。判別方式分為兩種：投票法及加權投票法，分類結果為類別最多的組別，兩者差別為後者考慮近鄰之距離遠近進行加權，再進行分類。優點是精度高且無資料輸入假定、資料型態不受限，缺點是時間空間複雜度高，樣本平衡度依賴高。

羅吉斯回歸

主要在探討依變數與自變數之間的關係，與一般線性回歸不同處在：依變數為類別變數，特別是分成兩類的變數。利用自變數的線性組合與連結函數預測發生步態凝凍的機率，透過連結函數將線性組合轉換為分布在 0 ~ 1 之間之機率值，若機率大於門檻值則判定會發生步態凝凍。優點是方便計算與訓練，且自變數不需常態分佈的假設，故限制較少。

* 降維

Relief f

是從維度縮減的 Relief 法延伸出的改良方法，相較於Relief法，能處理多類別問題，適用於各種型態的資料。其步驟是先給定所有變數的初始權重，透過不斷遞迴計算，最後挑選出權重較高的前幾名變數作為較為重要的代表。其遞迴過程是選定一個隨機樣本，計算與其同類和異類的曼哈頓距離，同類的距離越小代表越像此類則增加同類樣本權重，異類的距離越大則減少異類樣本權重，能較 Relief 法減少雜訊干擾，得到更穩健的結果。

互資訊 (Mutual Information, MI)

度量了兩個變數之間相互依賴的程度。對於兩個隨機變數，MI是在獲得一個隨機變數的資訊之後，觀察另一個隨機變數所獲得的「資訊量」。相互資訊的概念與隨機變數的熵緊密相關。MI愈大代表所降低的不確定性愈多，則關聯性愈強。MI不僅度量實值隨機變數和線性相關性(如相關係數)，更可捕捉兩變數間任何形式的相依程度。

最小冗餘最大特徵法 (minimum-redundancy maximum-relevance, mRMR)

以迭代方式工作，每次迭代都會透過計算分數來選出最佳特徵，被選入特徵集後就不再被挑出，而特徵的選擇與結果有高度相關，且特徵集內的特徵之間相關性最小，以降低冗餘性。類別型可利用互資訊(MI)計算，連續型可利用互資訊或F檢定及Pearson相關係數計算。此演算法只適用於監督式學習，且計算快速。

* 訓練集和測試集的切割

因凍凝步態(目標類)與非凍凝步態(非目標類)筆數懸殊，在所有的視窗中，目標類占不到20%，避免分類器產生偏誤，故先將資料平衡，取所有目標類視窗，再抽取目標類視窗數兩倍的非目標視窗。將目標類與非目標類各採70%合併為訓練集，測試集則為剩餘的各30%合併，整理成較為平衡的資料集。重複30次抽取非目標視窗，又因非目標類視窗遠超過目標類視窗，故每次抽取之視窗不一定相同，讓非目標類視窗之特徵能較完整的被訓練，共產生30個訓練集及30個測試集。

* 指標評估

上述所有統計分析方法皆運用排除第四位及第十位病患外的資料做 7:3 等比例的交叉驗證來檢視模型預測能力的好壞，其值越高代表預測能力越佳。此外，各分析結果將呈現如表一形式：真陽性 (True Positive, TP)代表為該視窗實際為凝凍步態且預測該視窗亦為凝凍步態；真陰性 (True Negative, TN)代表為該視窗實際為非凝凍步態且預測該視窗亦為非凝凍步態；偽陰性 (False Negative, FN)代表該視窗實際為凝凍步態且預測該視窗為非凝凍步態；偽陽性 (False Positive, FP)該視窗實際為非凝凍步態且預測該視窗為凝凍步態，並依此建構各項評估指標，評估統計方法的分類表現之優劣。各項評估指標陳列於表二。

表一 預測結果表示

表二 統計評估指標

|  |  |  |
| --- | --- | --- |
| 評估指標 | 說明 | 定義 |
| 召回率(Recall) | 真實狀態為陽性  其預測結果亦陽性的比率 |  |
| 特異度(Specificity) | 真實狀態為陰性  其預測結果亦為陰性的比率 |  |
| 準確率(Accuracy) | 測量整體分類準確率 |  |

|  |  |  |
| --- | --- | --- |
| 真實狀態 | 預測結果  陽性 陰性 | |
| 陽性  陰性 | TP  FP | FN  TN |