實作心得報告

NICE: an algorithm for nearest instance counterfactual explanations

簡述NICE:

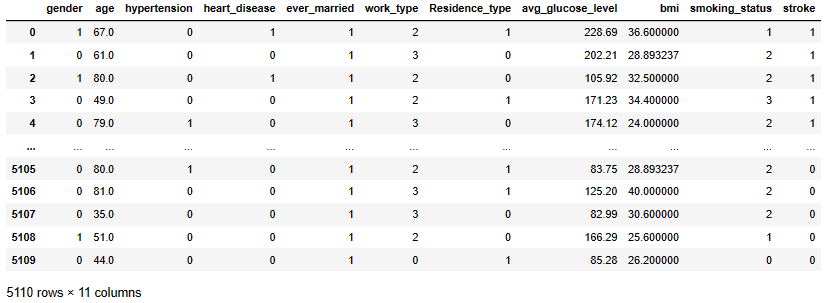
NICE演算法的核心目標在於提供對分類結果的反事實解釋，這意味著在保持其他features變動最少的情況下，揭示出將樣本分類至不同類別的可能性。這種能力不僅有助於使用者深入了解機器學習模型的判斷過程，更提供了一個實際的指引，讓使用者了解在什麼條件下可以改善或調整以達到更理想的結果，此演算法的特點在於因為高速，且保證找到counterfactual。

使用申請貸款舉例，要成功申請貸款，或許需要提高月收入這一屬性。

實作方法:

我們的實作方式，是使用其他dataset觀察輸出結果，我們使用了kaggle上面取得的資料集: [Stroke Prediction Dataset (kaggle.com)](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data) 。

此dataset展示了年齡、血糖、高血壓等與中風息息相關的徵兆，

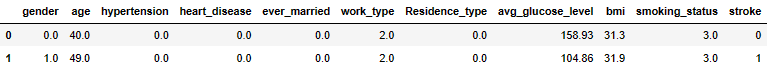


以及諸如工作，居住地點，婚姻狀態等隱性原因的類別資料。

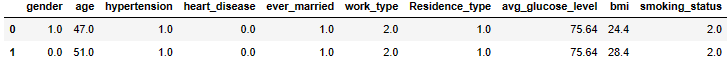
目前的NICE演算法不夠全面，所以只能使用二元分類結果的資料集，例如上述的預測有沒有中風。

而NICE演算法支援使用三種不同的unlike neighbor的方法，分別為Sparsity(只使用較少的features),Proximity(nearest unlike neighbor),Plausibility(結果需要符合現實邏輯)，本次使用的是最基本的proximity。

得到以下結果:



此結果中第0行為起點，而第一行為NICE找到的nearest unlike neighbor, 如此透過此例子可以看到，血糖濃度雖下降嚴重，但是年齡的提升能讓此患者患上中風，而兩個的性別也有所不同(0:male, 1:female)，在後續會做相關的討論，我們能透過此解釋概略地觀察模型，在此狀況中此模型將Age的權限提高了不少，

我們在觀察其他的instance。

得到的結果也與年齡較為相關，可以說年齡確實為主要因素，同時經由和一開始的例子做比較，可以看出年齡不管是男變女，還是女變男，都很難探討出對中風的影響，或許是個較無重要的因素。

心得:

經過實作，感覺到目前演算法還是有著不少的侷限性，但是其演算法非常明瞭簡潔，所以相信擴充性應該是非常良好的。