資四甲 110716035 吳宜昌

機器學習HW2\_Classification

程式碼語言：Python

資料集：HW2\_heart\_2020\_cleaned

內容：

1. **資料清理與視覺化圖表（程式碼Part1. 資料清理與視覺化圖表）：**

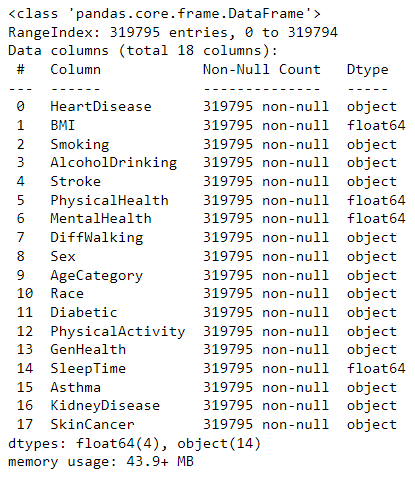


圖1檢查是否有NULL值

如上圖1，此資料並無任何Null值。

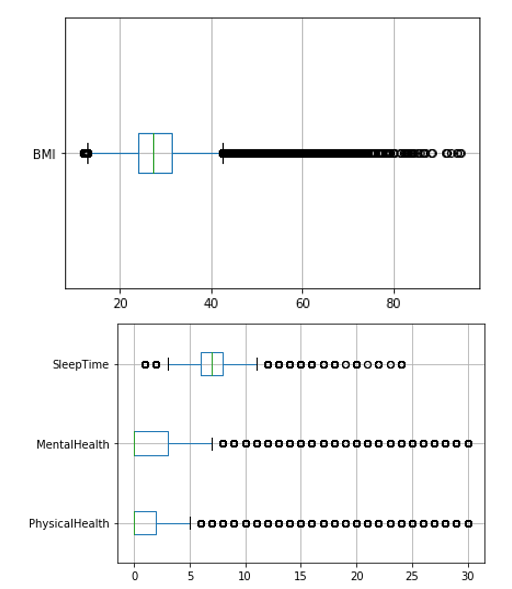
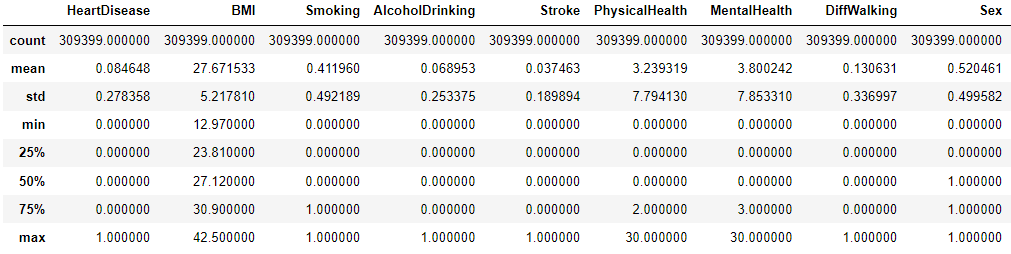


圖2資料集箱型圖

如上圖2，BMI的資料集因為過於分散，且太多離群值，因此我決定將BMI的離群值刪除。不處理PhysicalHealth與MentalHealth的原因是：經過測試之後，仍然無法有效刪除，因此此處不做處理。

1. **敘述性統計分析（程式碼Part2. 敘述性統計分析）：**



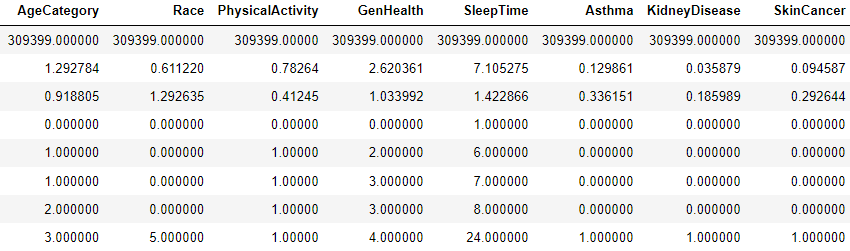


圖3敘述性統計分析表格

如上圖3，此為資料集heart\_2020\_cleaned的敘述性統計，包含所有類別的平均數、標準差、最小值、4分之1位數、中位數、4分之3位數、最大值。

1. **特徵相關性分析（程式碼Part3. 特徵相關性分析）：**

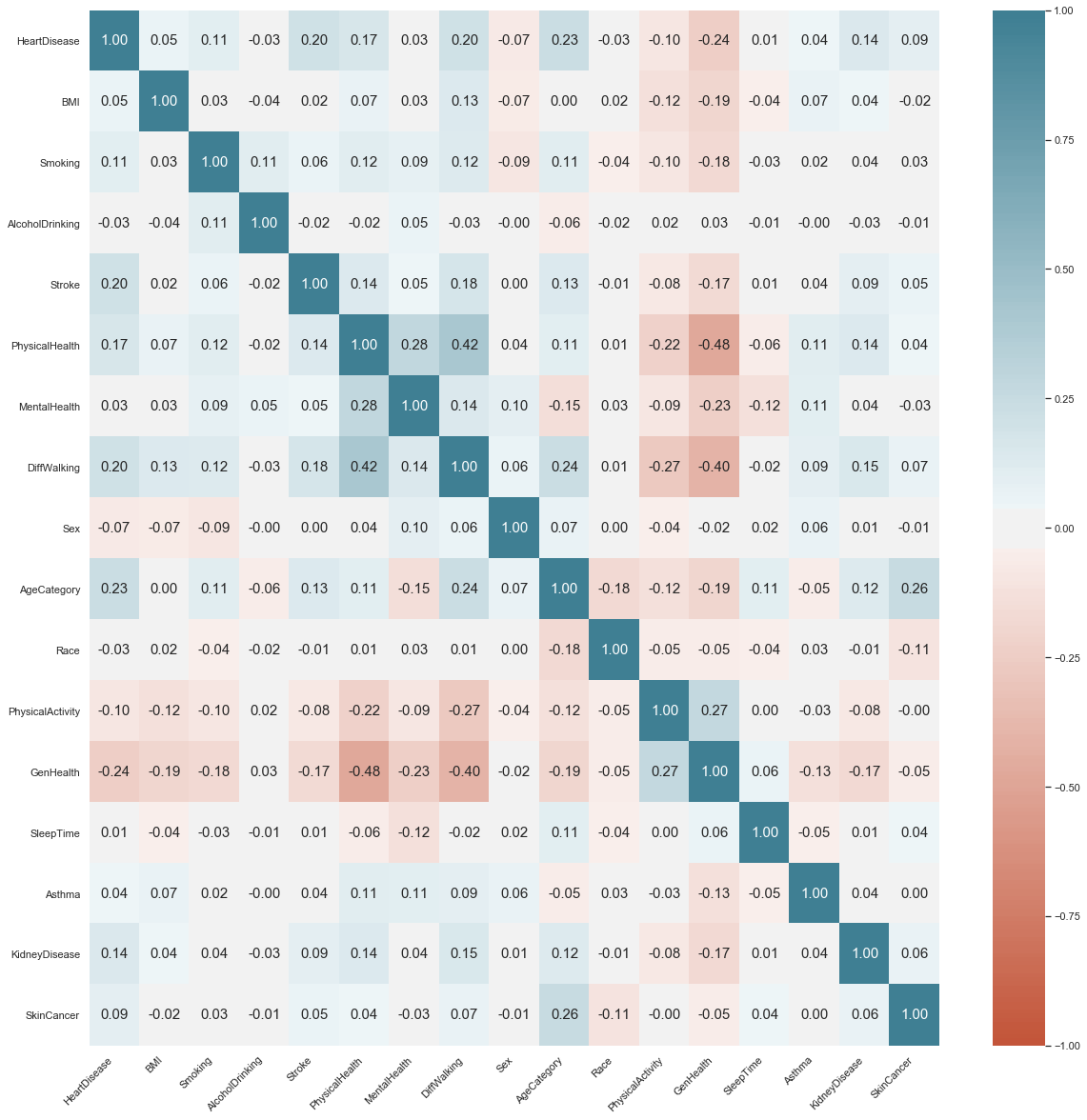


圖4熱力圖

如上圖4，此為資料集heart\_2020\_cleaned的特徵相關性分析。可從此圖得知與HeartDisease有正相關的較少，且並沒有出現中度或高度正相關的類別。因此我打算取與HeartDisease相關性相對比較高的3個類別出來，分別是Stroke、DiffWalking、AgeCategory。因為只有他們與HeartDisease的相關性係數有>0.2。

1. **資料分割與建置4個分類模型(1. Logistic regression、2. SVM、3. Random forest、4. KNN) （程式碼Part4. 資料分割與建置4個分類模型）：**

因此我將HeartDisease設為應變數，Stroke、DiffWalking、AgeCategory設為自變數，並將資料集以7：3的比例切割。

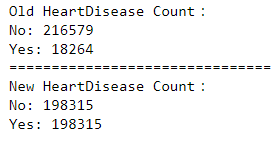


圖5回放式取樣

因為HeartDisease為Yes的資料集與No相比明顯過少，因此我將Yes的資料集進行回放式取樣。

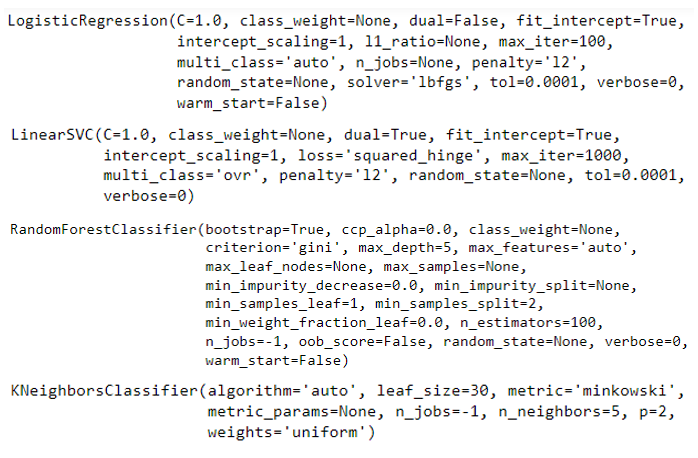


圖6模型建置

如上圖6，此為邏輯斯迴歸、SVM、隨機森林、KNN的模型建置。且SVM模型只迭代1000次進行訓練，隨機森林模型只以5層進行訓練。

1. **綜合比較4個模型的分類結果與分析討論（程式碼Part5. 綜合比較4個模型的分類結果與分析討論）：**

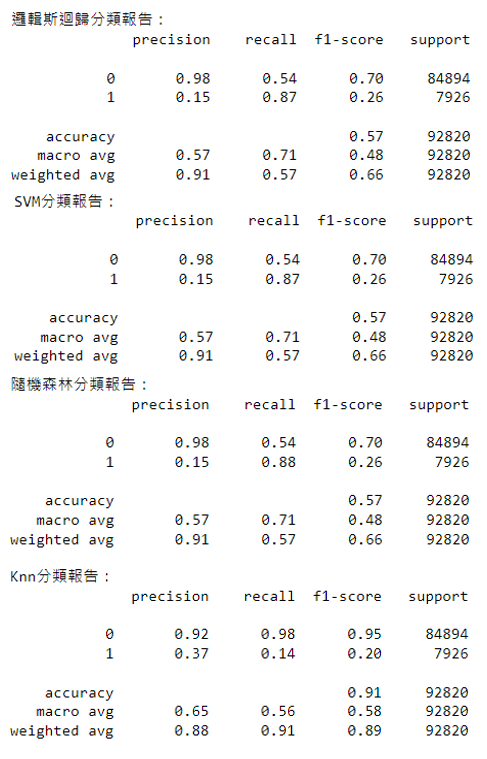


圖7分類報告

如上圖7，此為4種模型的分類報告，可明顯看得出來邏輯斯迴歸、SVM、隨機森林的Accuracy都差不多是0.57，但是KNN的Accuracy有0.91明顯比其他三者高。可以顯示KNN的效果可能是最佳的。



圖8交叉驗證

如上圖8，此為4種模型的交叉驗證的結果，很訝異的是雖然邏輯斯迴歸、SVM、隨機森林的Accuracy偏低，但是交叉驗證結果卻比KNN高。

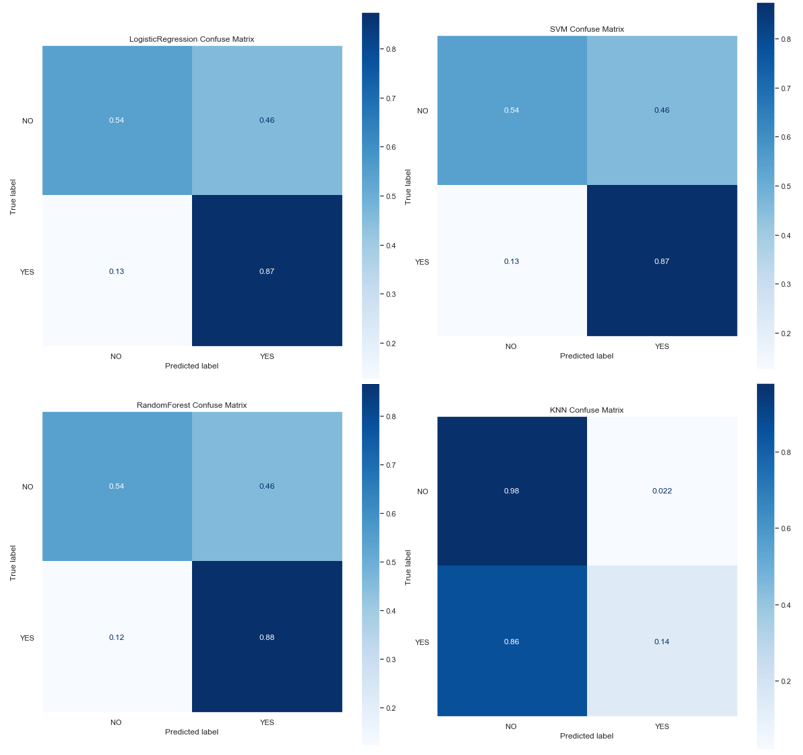


圖9混淆矩陣

如上圖9，此為4種模型的混淆矩陣，表4種模型在測試集的預測狀況。可以很明顯的看出來邏輯斯迴歸、SVM、隨機森林在預測HeartDisease是Yes的時候比較準，反而是預測是No的時候錯誤率比較高。且KNN在預測No的情況很準確，但是在預測是Yes的錯誤率卻偏高。

從此次結果觀察得知，雖然KNN模型的Accuracy很高，但是在預測方面只能較準確的預測No的情況，反而是Yes不太能準確預測到。

相對的邏輯斯迴歸、SVM、隨機森林此三種模型，雖然在Accuracy相對較低，但是能比KNN較準確的預測到Yes的情況，但是No還是不太能準確預測到。