2021 電子商務技術 作業六

第一部分：Predicting Performance

分類任務：根據一個人的相關資訊判斷此人 10 年內是否有罹患心臟病的風險 (TenYearCHD)。

1. 載入資料集 framingham_train.csv (全做為訓練用)，並建立 sklearn LogisticRegression 模型進行此分類任務，請將 solver 設為 lbfgs。(5%)

   (訓練資料 Accuracy：0.8523)

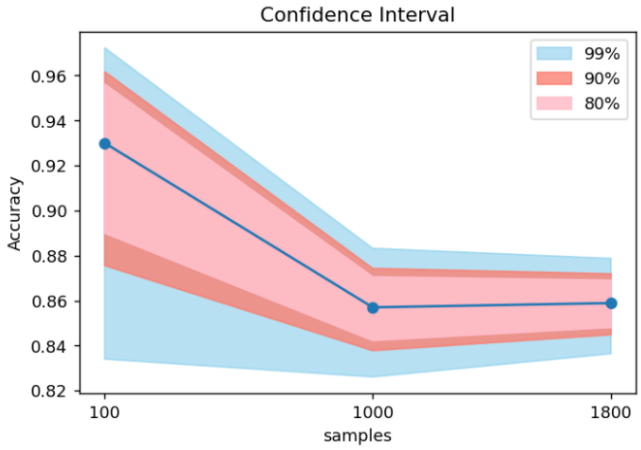
2. 載入資料集 test_sample_1.csv、test_sample_4.csv、test_sample_5.csv，並用程式算出模型在此 3 個測試資料集於信心水準 80%、90%、99%下的準確率信賴區間，輸出如下的表格。(分成兩個表格也可以) (20%)

   * 註：請在程式中自行建立 normal distribution 的 confidence limits 對照表，z 值取小數點後兩位。(Ch5, p16 Table 5.1)

樣本數

		80%	90%	99%
100	lower	0.889941	0.875531	0.834004
	upper	0.956196	0.961676	0.972323
1000	lower	0.842246	0.837764	0.826075
	upper	0.870586	0.874298	0.883204
1800	lower	0.848059	0.844807	0.836394
	upper	0.869066	0.871887	0.878739

3. 視覺化上題的信賴區間，X 軸為樣本數、Y 軸為準確率，且須有 legend，如下圖所示。(只畫線也可以，但要能看出每條線代表什麼) (20%)



4. (a)在同一信心水準下， 3 個測試資料集的信賴區間有何不同？(b) 不同信心水準下的信賴區間有何不同？(15%)

第二部分：Model Comparison

1. 載入資料集 BreastCancer.csv，使用 10-fold cross validation 評估 Naïve Bayes、Decision Tree 和 SVM 這三種演算法的表現(Accuracy、Precision、Recall)。請印出模型 10 次的測試結果和平均，並比較各模型的表現。(40%)

   * 註：資料前處理、模型參數等請自由操作。

- **繳交期限：5/5(三) 中午 12:00**
- **請繳交答案卷(.pdf)和程式檔(.ipynb)，檔名為 ECT_HW6_學號。pdf 檔內註明題號，將該小題的程式及執行結果截圖貼上，如題目有詢問額外的問題也請一併寫上**
- **上傳至 ee-class 作業區，遲交一天扣該次作業 5%**