2021 電子商務技術 作業四

第一部分：Linear Regression

使用資料集 retail_transactions.csv，利用顧客在 2010 的消費資料預測其 2011 年的消費。請使用 Python 完成下列題目。

1. 刪除有空值的資料以及根據 2010/2011 revenue 刪除 outliers，此處 outliers 的定義為大於中位數 3 個標準差。(請在刪除 outlier 前先算好中位數、標準差)(3%)

2. 計算各屬性與 target output (2011 revenue)的相關係數(2%)，請寫出與 output 最相關和最不相關的屬性(2%)，並刪除與 output 無關的屬性(相關係數< |0.1|)(2%)。

3. 使用 sklearn 的 train_test_split(random_state=15)函數將資料分為訓練集與測試集，比例為 67%：33%。(2%)

4. 使用 sklearn 及訓練集建立 LinearRegression 模型。(15%)

5. 印出模型中的係數和截距(5%)，並還原此 linear equation(2%)。

6. (a) 預測測試集中的顧客在 2011 年的消費(5%)，(b) 印出模型在測試資料的 RMSE(root mean squared error)和 MAE(mean absolute error)(3%)。

(** RMSE = 3682.95, MAE = 2174.62) (using default model parameters)

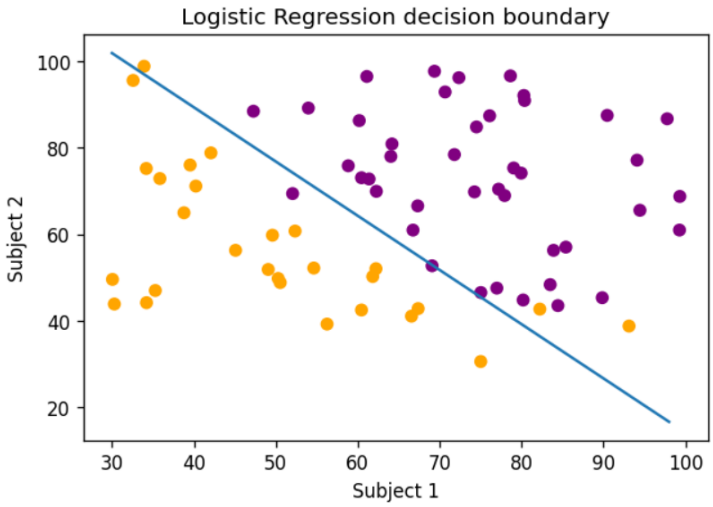
第二部分：Logistic Regression

使用資料集 exam_score.csv，根據學生某兩科的成績預測其這學期是否能及格(0 為不及格，1 為及格)。請使用 Python 完成下列題目。

7. 使用 sklearn 的 train_test_split(random_state=1)將資料分成訓練集與測試集，比例為(7:3) (2%)。

8. (a) 使用 sklearn 及訓練集建立 LogisticRegression 模型(15%)，(b) 印出模型對訓練資料的準確率(Accuracy) (3%)。

9. 請視覺化此模型對訓練資料的分類，如下圖所示。(圖中須能區分出不同類型的資料點) (10%)



10. (a) 使用模型對測試集做預測(2%)，(b)印出模型預測的準確率(Accuracy) (2%)。

11. 有一個新學生兩個科目的分數分別為 45、80 分，請預測此學生這學期是否能及格(3%)，模型預測此學生能及格的機率為多少(2%)。

(** Accuracy on training data = 0.93, Accuracy on test data = 0.87) (using default model parameters)

請使用 Weka 完成下列題目

12. 使用 Logistic Regression 對 exam_score.csv 進行分類，並使用所有資料做訓練。(15%)

13. 請比較 weka 與 python sklearn 建立的 Logistic Regression 模型結果的差異。(5%)

- **繳交期限：4/14(三) 中午 12:00**
- **請繳交答案卷(.pdf)和程式檔(.ipynb)，檔名為 ECT_HW4_學號**
- **上傳至 ee-class 作業區，遲交一天扣該次作業 5%**