統計學習與資料探勘期中考試 2025/11/5

學號： _____
姓名： _____

1. 使用**"Data1"** 資料集進行迴歸分析

   請使用**"Data1"** 資料集，其中應變數為y，自變數為$V_1, V_2, V_3, V_4, \ldots, V_{10}$。請回答下列問題：

   - 當使用具有隨機性的演算法（例如隨機抽樣）時，請設定`random_state=2025`以確保結果可重現。
   - 若所選模型**不具尺度不變性（scale-invariance）**，請在建模前對輸入的自變數進行標準化（standardization）。

   (a) 繪製所有變數的成對散佈圖（pair plot）。你觀察到什麼現象？是否有變數與y 呈現線性關係（linear relationship）？（6%）

   (b) 使用`train_test_split` 隨機選取一半的觀測資料作為訓練集（training set），其餘一半作為測試集（test set）。（5%）

   (c) 使用訓練集，先以V_{10} 為唯一自變數擬合一個簡單線性迴歸模型（simple linear regression model），再使用所有變數擬合一個多元線性迴歸模型（multiple linear regression model）。根據V_{10} 在這兩個模型中的p 值，你認為V_{10} 是一個顯著的預測變數嗎？（10%）

   (d) 計算所有自變數的VIF（Variance Inflation Factor）。接著，將V_{10} 從模型中移除，並對剩下的九個變數進行多元線性迴歸。請描述你觀察到什麼現象？（10%）

   (e) 使用訓練集，透過逐步向後選取法（backward stepwise selection），根據貝葉資訊準則（BIC，Bayesian Information Criterion）找出最佳線性迴歸模型。此方法選出了哪些變數？（10%）

   (f) 在訓練集上擬合一個Ridge regression 模型，並繪製十個預測變數在標準化後的迴歸係數（standardized coefficient estimates）隨著λ（或α）變化的圖形。找出能使10 折交叉驗證誤差（10-fold cross-validation error）最小的λ（或α）數值。在進行交叉驗證與繪圖時，請從5×10^4 掃描到5×10^{-4}，並以對數刻度均勻取100 個樣本點。請回報對應的訓練均方誤差（training MSE）與測試均方誤差（test MSE）。最後請回答是否有估計出的迴歸係數為0？（15%）

   (g) 在訓練集上擬合偏最小平方法迴歸模型（PLS，Partial Least Squares regression model），並透過留一交叉驗證（LOOCV，Leave-One-Out Cross-Validation）選擇成分數M。請回報所選擇的M 值與對應的訓練均方誤差與測試均方誤差。（10%）

2. 使用**"Data2"** 資料集進行分類分析

請使用**"Data2"** 資料集，其中應變數為y，自變數為$V_1, V_2, \ldots, V_{15}$。請回答下列問題，並使用**加權F1 分數（weighted F1-score）**作為模型效能的評估指標。

- 可參考`sklearn.metrics` 模組中的方法計算F1 分數，其定義為：

$$\text{F1} = \frac{2 \cdot (\text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}}$$

- 加權F1 分數（weighted F1-score）的定義如下：

$$F1_{\text{weighted}} = \sum_{i=1}^{K} \left(\frac{N_i}{N} \cdot F1_i \right)$$

  其中：

  - K：類別總數；
  - $F1_i$：第i 類別的F1 分數；
  - N_i：第i 類別的真實樣本數（support）；
  - $N = \sum_{j=1}^{K} N_j$ 為總樣本數。

- 當使用具有隨機性的演算法時，請設定`random_state=2025` 以確保結果可重現。
- 若所選模型**不具尺度不變性（scale-invariance）**，請在建模前對輸入的自變數進行標準化處理。

(a) 將資料集以0.3 的測試集比例分割為訓練集與測試集。分別統計類別0 與類別1 的樣本數。此資料集是否為平衡資料（balanced dataset）？（5%）

(b) 使用訓練集建立邏輯斯迴歸模型（logistic regression model），並回答訓練集與測試集上的加權F1 分數。（8%）

(c) 將分類的門檻值（threshold）從0.5 調整為0.25，並回報訓練集與測試集上的加權F1 分數。與(b) 小題相比，你觀察到什麼變化？（8%）

(d) 使用訓練集建立線性判別分析分類器（LDA，Linear Discriminant Analysis），並回報訓練集與測試集上的加權F1 分數。（8%）

(e) 使用訓練集建立樸素貝氏分類器（Naive Bayes classifier），並回報訓練集與測試集上的加權F1 分數。（8%）

(f) 解釋過擬合（overfitting）與欠擬合（underfitting）的意義。我們應該如何根據訓練集（training set）、驗證集（validation set）與測試集（test set）來選擇模型？（7%）