112 年度第 2 次 巨量資料分析師能力鑑定 初級試題

科目 2：B12 資料處理與分析概論　　卷號：B12-2201

考試日期： 112 年 11 月 25 日　　第 1 頁，共 16 頁

單選題 50 題（佔 100%）

A	1. 當需要從大量的文字資料中提取電話號碼，但只有一次機會進行掃描，下列哪一種策略是最有效的？ (A) 先使用字串切割方法，再利用正則表達式匹配 (B) 單獨使用正則表達式進行全文匹配 (C) 利用 NLP 工具先識別文字語境再匹配 (D) 透過簡單的字串比對找到匹配項
D	2. 在正則表達式中，下列哪一項「不」是常見的匹配模式？ (A) 字元 (B) 數字 (C) 日期 (D) 邏輯運算符
D	3. 填補數值資料空值時，下列哪一項「不」是較適合的選項？ (A) 平均數 (B) 中位數 (C) 眾數 (D) 最大值
C	4. 資料縮減（Data reduction）「不」包含下列哪一種方式？ (A) 降維（Dimensionality reduction） (B) 減少數量（Numerosity reduction） (C) 資料加密（Data encryption） (D) 資料壓縮（Data compression）
A	5. 假設您每分鐘都會收到某張股票的開盤價、收盤價、最低價、最高價、成交量，若您只想儲存收盤價，最適合 R 語言中的哪一種結構？ (A) 向量（Vector） (B) 矩陣（Matrix） (C) 字串（Character） (D) 資料框架（Data frame）
D	6. R 語言中，下列哪一項是專門在處理群組與摘要的函數？

	(A) spread() (B) sort() (C) gather() (D) aggregate()
B	7. 數值變數最常用的分散程度是變異（Variance），關於變異數的敘述，下列哪一項正確？ (A) 標準差（Standard deviation）之開方根為變異數 (B) 變異係數（Coefficient of variation）是標準差與平均數的比值 (C) 中位數絕對離差（Median Absolute Deviation）是計算各觀察值與平均值的距離值後，再取其中位數 (D) 變異數、標準差、變異係數等統計變異量數，適用於類別變數
D	8. 以單變量來說，預測變數只有一個獨一無二的值，稱為分佈退化（Degenerate Distribution），對巨量資料分析模型是沒有貢獻的。下列哪一項「不」是判斷分佈退化的方法？ (A) 從統計量數中的變異數辨識，判斷量化變數是否為零變異 (B) 從統計量數中的變異數辨識，判斷量化變數是否為近乎零變異 (C) 由直方圖或密度曲線，觀察有無極度右偏、極度左偏或單一高峽峰的現象 (D) 檢查變數中有沒有空值（Null）
D	9. 巨量資料下特徵選取（Feature selection）的工作十分重要，下列何項屬於非監督式特徵過濾（Unsupervised filtering）方法？ (A) 運用卡方檢定（Chi-squared test） (B) 計算訊息增益（Information gain） (C) 計算費雪分數（Fisher score） (D) 計算相關係數（correlation coefficient）
C	10. 如附圖所示，是個人行動電話服務記錄表（只顯示前五筆），

	如果想用 Python 產生各種服務類型（item）下使用時間（duration）的群組與摘要報表，下列哪一項指令最「不」可能運用到？
	`##  date duration item month network`
	`## 0 15/10/14 06:58 34.429 data 2014-11 data`
	`## 1 15/10/14 06:58 13.000 call 2014-11 Vodafone`
	`## 2 15/10/14 14:46 23.000 call 2014-11 Meteor`
	`## 3 15/10/14 14:48 4.000 call 2014-11 Tesco`
	`## 4 15/10/14 17:27 4.000 call 2014-11 Tesco`
	`## network_type`
	`## 0 data`
	`## 1 mobile`
	`## 2 mobile`
	`## 3 mobile`
	`## 4 mobile`
	(A) AGG
	(B) SUM
	(C) TYPE
	(D) GROUP BY

A	11. 如附圖所示，在 Python3 中以 Pandas 進行資料的操作。附圖表格一（df1）、表格二（df2）為拉拉五金行的訂單資料，其中 TransactionId 代表各個銷售地點所發生的交易記錄編號。若想將表格一、表格二合併，得到如表格三之執行結果（需符合圖中欄位順序），可以執行下列哪一個選項中的程式碼？

[表格一：df1]

	TransactionId	銷售地點	訂單時間
0	1	台北門市	2022-01-01 10:13:22
1	1	高雄門市	2022-01-01 11:14:27
2	2	台北門市	2022-01-01 11:17:09
3	1	網路出貨	2022-01-01 12:02:43
4	2	網路出貨	2022-01-01 12:05:19
5	3	網路出貨	2022-01-01 12:26:00
6	2	高雄門市	2022-01-01 12:47:38
7	3	台北門市	2022-01-01 13:30:56

[表格二：df2]

	TransactionId	銷售地點	UserId	商品代號	訂單金額	訂單數量
0	1	台北門市	101	A	100	1
1	1	高雄門市	102	B	200	1
2	2	台北門市	103	C	300	1
3	1	網路出貨	104	D	400	1
4	2	網路出貨	105	B	200	1
5	3	網路出貨	106	A	100	1
6	2	高雄門市	107	C	300	1
7	3	台北門市	108	C	300	1

[表格三]

	TransactionId	銷售地點	訂單時間	UserId	商品代號	訂單金額	訂單數量
0	1	台北門市	2022-01-01 10:13:22	101	A	100	1
1	1	高雄門市	2022-01-01 11:14:27	102	B	200	1
2	2	台北門市	2022-01-01 11:17:09	103	C	300	1
3	1	網路出貨	2022-01-01 12:02:43	104	D	400	1
4	2	網路出貨	2022-01-01 12:05:19	105	B	200	1
5	3	網路出貨	2022-01-01 12:26:00	106	A	100	1
6	2	高雄門市	2022-01-01 12:47:38	107	C	300	1
7	3	台北門市	2022-01-01 13:30:56	108	C	300	1

(A) pd.merge(df1, df2, on=['TransactionId', '銷售地點'], how='left')

(B) pd.merge(df1, df2, on='TransactionId', how='left')

(C) pd.merge(df2, df1, on=['TransactionId', '銷售地點'], how='right')

(D) pd.merge(df2, df1, on='TransactionId', how='right')

D

12. 假設有兩個表格資料（pandas.DataFrame 格式）df1 和 df2，分別包含 id 和 value 欄位。若要將 df1 和 df2 依照 id 欄位進行合併，並保留兩個表格的所有資料（包括那些只出現在一個表格的資料），請問下列程式碼哪一項正確？

(A) pd.merge(df1, df2, on='value', how='outer')

(B) pd.merge(df1, df2, on='value', how='inner')

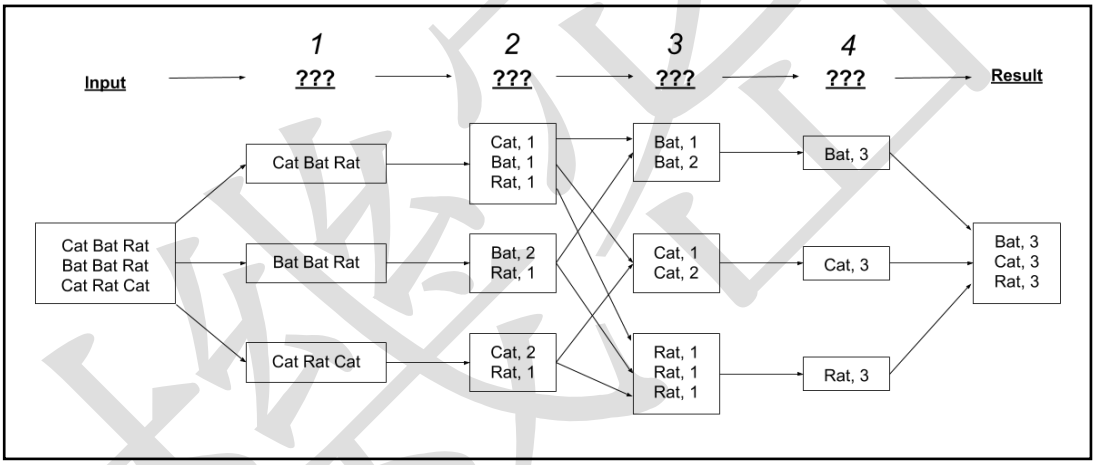
(C) pd.merge(df1, df2, on='id', how='inner')

(D) pd.merge(df1, df2, on='id', how='outer')

C

13. 關於屬性萃取（Feature Extraction）與屬性挑選（Feature Selection），下列敘述哪一項錯誤？

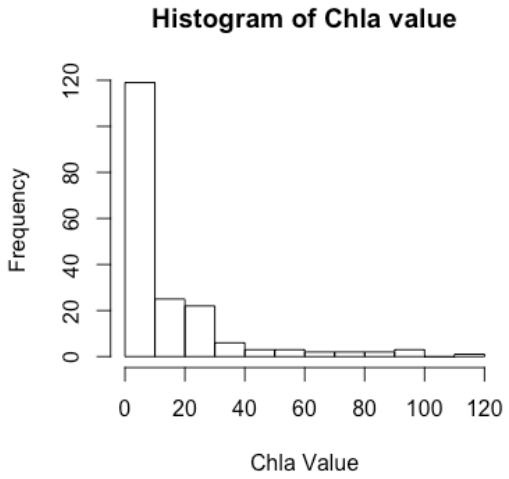
	(A) 主成分分析（principle Components Analysis, PCA）是一種屬性萃取的技術 (B) 屬性萃取可以減少屬性間的相互影響，降低模型的複雜度 (C) 屬性轉換中移除分佈異常的屬性，不是屬性挑選的方法 (D) 屬性挑選與屬性萃取皆可用於降低資料維度
A 或 D	14. 在電子商務資料庫中，有一個類別型資料欄位 "產品類別"，包含了 "衣服"、"家電"、"手機" 等多個類別。請問在 Python 中使用哪個套件及函式，可將這個文字類別資料欄位轉換為數值類別資料欄位？ (A) pandas 的 get_dummies() (B) pandas 的 astype() (C) numpy 的 where() (D) sklearn 的 LabelEncoder()
A	15. 考慮一個包含三維資料的資料集，想要將其降維至二維，同時盡可能保持在三維空間中的資料特性。下面哪一個方法在實現這一目標時效果較佳？ (A) 等距特徵映射（Isometric Feature Mapping, Isomap） (B) 主成分分析（Principle Component Analysis） (C) 線性判別分析（Linear Discriminant Analysis） (D) 獨立成分分析（Independent Component Analysis）
A	16. 如附圖所示，MapReduce 程式設計的邏輯流程，一般包含下列步驟，下列哪一項是正確的順序？ a. Input b. output c. map d. reduce e. split f. shuffle & sort

	(A) a -> e -> c -> f -> d -> b (B) a -> e -> c -> d -> f -> b (C) a -> c -> e -> d -> f -> b (D) a -> c -> d -> e -> f -> b
D	17. 下列哪一個「不」是關於 HDFS 資料儲存的特色？ (A) 使用元資料（Metadata）來描述檔案的屬性 (B) 使用分散式儲存來儲存資料 (C) 使用資料區塊（Data Block）來儲存資料 (D) 使用檔案系統來管理資料
B	18. 如附圖所示，藉由 MapReduce 進行詞頻統計（word count）工作的流程示意圖。請問圖中對應的動作組合應為下列哪一項？  (A) Mapping -> Splitting -> Shuffling -> Reducing (B) Splitting -> Mapping -> Shuffling -> Reducing (C) Reducing -> Splitting -> Mapping -> Shuffling (D) Splitting -> Shuffling -> Mapping -> Reducing
D	19. 一家英文新聞社正在使用自然語言處理技術對他們的新聞報導進行情感分析。在進行分析之前，下列哪一項「不」是適當的文字前處理步驟？ (A) 將文本轉換為小寫 (B) 刪除停用詞（Stop Words） (C) 進行詞幹提取，將 "running"、"runner" 都轉換為 "run"

	(D) 使用 Word2Vec 對文字進行情感標記
A	20. 關於 MapReduce 架構，下列敘述哪一項錯誤？ (A) 當節點（Nodes）失效時，其他節點無法接管失效的任務 (B) 可運行在不可靠的低階電腦叢集（Clusters）上 (C) 映射（Map）是將任務分配到不同節點進行計算 (D) 化簡（Reduce）是將處理完的結果重新組合
C	21. 如附圖所示的亂數表，班上有 50 位學生，座號為 01 至 50。今要找 5 位同學參加比賽，經討論後決定以簡單隨機抽樣法選取人選，若以亂數表之第一列第一行為起始點由左至右選取，則會抽到哪些座號，下列哪一項正確？ 2 6 4 2 8 4 1 4 9 3 2 5 7 9 0 8 9 8 3 7 5 7 1 7 6 7 8 8 0 7 0 7 9 5 6 2 3 9 0 3 9 8 4 3 5 9 9 4 6 6 1 8 6 9 4 5 4 9 3 3 9 0 9 5 8 0 8 0 5 9 5 0 6 4 8 5 5 2 5 8 2 7 0 9 4 6 7 2 7 4 6 1 1 1 0 8 6 4 6 2 (A) 02, 06, 04, 28, 41 (B) 26, 42, 04, 14, 32 (C) 26, 42, 14, 25, 08 (D) 26, 42, 41, 49, 32
B	22. 關於共變異數（Covariance）與相關係數（Correlation Coefficient）的敘述，下列哪一項錯誤？ (A) 共變異數其值的正負號代表兩隨機變數是正向或負向關聯 (B) 共變異數其值不僅能看出兩變數相關的方向，也能看出兩變數間關係的強度 (C) 相關係數其值介於 -1 到 1 之間 (D) 相關係數其值是將共變異數標準化
D	23. 關於抽樣分配的敘述，下列哪一項錯誤？ (A) 母體為常態分配，抽樣平均數分配也會是常態分配

	(B) 抽樣分配的平均數估計不會隨著樣本而改變 (C) 母體不是常態分配，抽樣夠多時抽樣平均數分配會接近常態分配 (D) 抽樣分配平均數的區間估計，會隨著樣本變大而變大
D	24. 關於假設檢定的敘述，下列哪一項錯誤？ (A) H_0 為真，拒絕 H_0 是型 I 錯誤（Type I Error） (B) H_0 為真，沒有拒絕 H_0 的機率為 $1-\alpha$ (C) H_0 為假，沒有拒絕 H_0 是型 II 錯誤（Type II Error） (D) H_0 為假，拒絕 H_0 的機率為 α
C	25. 運用程式統計分析中，下列哪一個選項最能描述「中位數」的概念？ (A) 資料集中的最常出現的值 (B) 資料集中所有值的算術平均值 (C) 將資料集分為兩個相等比例 (D) 資料分布的對稱中心點
D	26. 設計程式分析兩個無序類別變數間的關係時，應使用下列哪一種統計方法？ (A) 皮爾森相關係係數（Pearson product-moment correlation coefficient） (B) 斯皮爾曼等級相關係數（Spearman's rank correlation coefficient） (C) 線性迴歸分析（Regression Analysis） (D) 費雪精確檢定（Fisher's exact test）
C	27. 在統計檢定中，「虛無假設」通常表示下列哪一項？ (A) 研究假設成立 (B) 研究假設不成立 (C) 母體參數間無顯著差異 (D) 母體參數間存在顯著差異

B	28. 在預測變數空間中，採行集群、關聯、降維等分析與探究手法，這種學習方式稱為下列哪一項？ (A) 監督式學習（Supervised learning） (B) 非監督式學習（Unsupervised learning） (C) 集成學習（Ensemble learning） (D) 強化式學習（Reinforcement learning）
D	29. 關於在統計機器學習建模的過程中產生的模型配適說明，下列哪一項錯誤？ (A) 配適不足是指測試與訓練誤差均高 (B) 過度配適是指測試誤差高，訓練誤差低 (C) 配適良好是指測試誤差低 (D) 配適狀況不明是指測試誤差高，訓練誤差低
D	30. 混淆矩陣（Confusion Matrix）的行列交叉會得到 (1) 真陽數（True Positive, TP）、(2) 真陰數（True Negative, TN）、(3) 假陽數（False Posotive, FP）、(4) 假陰數（False Negative, FN）。下列哪一個指標的公式錯誤？ (A) 正確率 accuracy = (TP+TN) / (TP+FN+FP+TN) (B) 敏感度 sensitivity = TP / (TP+FN) (C) 精確度 precision = TP / (TP+FP) (D) 假陽率 False Positive Rate, FPR = FP / (TP + FP)
D	31. 如附圖所示為海藻資料集的變數 Chla 的直方圖，請問其平均數與中位數最可能為哪一項？



(A) Mean: 5.5；Median: 14

(B) Mean: 5.5；Median: 60

(C) Mean: 14；Median: 14

(D) Mean: 14；Median: 5.5

D　32. 一家電子商務公司正在嘗試調校他們的推薦系統模型。他們使用了 k 折（K-Fold）交叉驗證法來確定模型的最佳超參數。在這種情境中，k 折交叉驗證的主要目的是什麼？

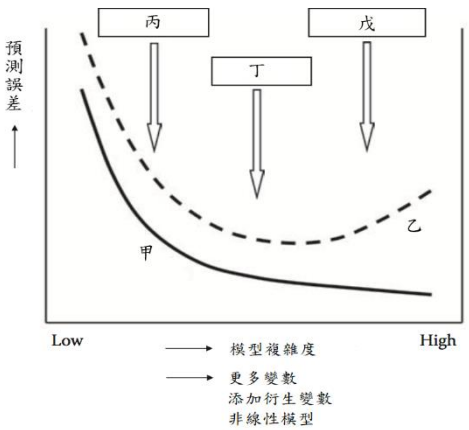
(A) 減少模型的計算時間

(B) 增加訓練數據的多樣性，以改進模型的泛化能力

(C) 減少模型的隨機誤差

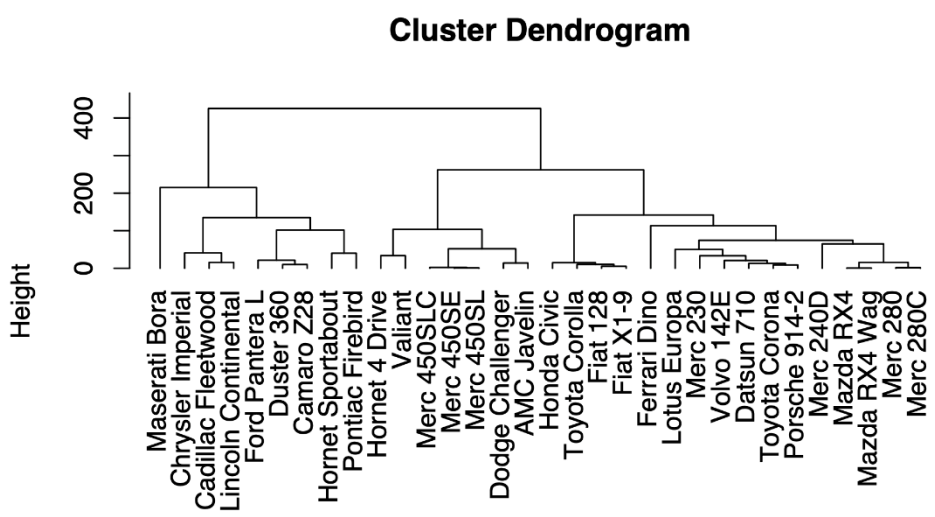
(D) 在多個子集上估計模型的性能，以避免過度配適

A　33. 如附圖所示為模型複雜度（Model Complexity）與預測誤差（Prediction Error）之間的變化關係，下列敘述哪一項正確？



	(A) 戊段表過度配適（Overfitting），它代表模型越複雜時與訓練集配適的過好，但卻逐漸喪失對測試集的預測能力 (B) 實曲線甲為測試集（Test set）樣本下的模型複雜度與預測誤差之間的變化關係 (C) 虛曲線乙為訓練集（Training set）樣本下的模型複雜度與預測誤差之間的變化關係 (D) 丙段表配適不足（Underfitting），此時訓練集預測誤差表現不佳，而測試集預測誤差表現良好
C	34. 許多領域中建立反應變數 y 與多個預測變數$x_1, x_2, \ldots, x_m$之間的模型，或稱關係$y = f(x_1, x_2, \ldots, x_m)$是一項基本的任務，其中 y 是系統中我們感興趣的事實性質，它最「不」可能被稱為下列哪一種選項？ (A) 目標（Target）變數 (B) 結果（Outcome）變數 (C) 解釋（Explanatory）變數 (D) 輸出（Output）變數
C	35. 下列哪一個選項最適合計算高維資料點之間的距離或相似性？ (A) 曼哈頓市街距離（Manhattan city block distance） (B) 歐幾里德直線距離（Euclidean distance） (C) 餘弦相似度（Cosine similarity） (D) 馬氏距離（Mahalanobis distance）
B	36. 在做資料探索時，經常會採用盒鬚圖（Boxplot）行離群值檢測，有關使用方式的敘述，下列哪一項錯誤？ (A) 盒鬚圖主要是由五個數值組成 (B) 盒鬚圖的中間數字為平均數 (C) 超出盒鬚圖範圍就可能被判定為離群值 (D) 盒鬚圖的上下屆範圍，是由第一四分位數減 1.5 倍

	四分位距（Interquartile range, IQR）以及第三四分位數加 1.5 倍四分位距所組成
C	37. 關於集群分析的敘述，下列哪一項錯誤？ (A) 由上而下的分裂法是階層式集群的一種 (B) 由下而上的聚合法是階層式集群的一種 (C) 空間密度集群法（Density-Based Spatial Clustering of Applications with Noise, DBSCAN）是以圖形為基礎的集群方法 (D) 輪廓係數（Silhouette coefficient）可評估各個樣本點歸群結果的優劣
C	38. 某醫院使用階層式集群分析對病人的體徵數據進行分類，目的是了解病人之間的相似性並優化治療方案。醫院使用下列哪一種圖表來可視化這種分析結果？ (A) 折線圖 (B) 散點圖 (C) 樹狀圖 (D) 柱狀圖
B	39. 某行動遊戲公司收集了玩家的遊戲時長、內購金額、等級等資料。該公司想要透過這些資料來找出哪些玩家更可能在未來進行內購。在進行資料探索時，應該使用什麼方式來評估「遊戲時長」和「內購金額」之間的關聯性？ (A) 計算兩者的共變異數 (B) 計算兩者的相關係數 (C) 進行 K-means 集群分析 (D) 使用階層式集群分析
D	40. 如附圖所示，下列敘述哪一項錯誤？



(A) 縱軸代表距離

(B) 橫軸代表樣本名稱

(C) 此圖名稱為樹狀圖（Dendrogram）

(D) 呈現 k 平均數集群（k-means clustering）的計算結果

C　41. 關於 k 平均數集群（K-means Clustering）的敘述，下列哪一項錯誤？

(A) 群數如果人為設定不佳，可能造成歸群結果不好

(B) 計算方法是初始化群中心後，指派各樣本的歸群隸屬後再更新各群中心座標，如此不斷循環直到結果穩定

(C) 無論樣本數的大小，均可運用 k 平均數集群的算法

(D) 不適合非球形、數據密度變化大或有離群數據的集群問題

D　42. 資料視覺化是貫穿資料探勘整個程序的重要技術，下列哪一項不是視覺化圖形常被使用的目的？

(A) 找尋數據中性質

(B) 找尋數據中的規律

(C) 找尋數據中的關係

(D) 找尋數據中的中位數

D　43. 關於監督式學習（Supervised Learning）的敘述，下列哪一項正確？

	(A) 訓練資料沒有標準答案，不需要事先以人力給予標籤（Label） (B) 機器在學習時，並不知道其分類是否正確 (C) 集群演算法（Clustering）為監督式學習方法之一 (D) 監督式學習是在訓練的過程中告訴機器答案，資料事先給標籤
D	44. 關於迴歸模型的敘述，下列哪一項正確？ (A) 共線性檢測是為了檢定自變數和因變數之間的相關性 (B) Durbin-Watson 是用來檢定自動相關，其範圍介於 0 ~ 4 之間，若有自動相關 DW 統計值會接近於 4 (C) 虛擬變數（Dummy Variable）通常用來表示一個類別變數（Categorical Variable），且不限於二分法 (D) 因變數 Y 的資料分配是二項式（Binary）應使用羅吉斯迴歸（Logistic Regression）
D	45. 關於羅吉斯迴歸模型的敘述，下列哪一項正確？ (A) 羅吉斯迴歸模型不屬於廣義線性模型（Generalized Linear Models, GLM） (B) 羅吉斯迴歸是建立多元類別機率值之成功勝率對數值的線性模型 (C) 羅吉斯迴歸模型不是預測事件機率，而是預測該事件是否發生 (D) 醫學檢驗中的陽性反應、貸款的違約事件都可用羅吉斯迴歸模型
D	46. 如附圖所示有一迴歸分析，請問下列哪一項正確？

迴歸方程式 $y = -1.41 + 0.0235x_1 + 0.00486x_2$

分析表

變數	係數估計	標準誤	t 值
常數(截距)	-1.4053	0.4848	
X_1	(A)	0.008666	
X_2	0.00486	0.001077	(B)

來源	d.f.	SS	MS	F
迴歸	2	1.76209	0.881045	52.30616
殘差	7		0.016844	
總計	9	1.88000		

R^2 = (C)

調整後的 R^2 = (D)

(A) 0.5

(B) 0.00486

(C) 0.881045

(D) 0.91936

B　47. 使用線性迴歸模型時，下列哪一種評估指標，可用來衡量模型的好壞？

(A) 輪廓係數（Silhouette Coefficient）

(B) 調整後的 R 平方值

(C) 混淆矩陣

(D) AUC-ROC 曲線

A　48. 相對於非監督式學習，下列哪一個是監督式學習方法的缺點？

(A) 一定要有資料標籤才能夠執行

(B) 訓練過程中有明確的正確答案可以供模型驗證

(C) 能夠較輕易的評斷模型的優劣

(D) 可以針對自己想要的應用場景進行標籤調整

D　49. 關於判定係數（Coefficient of Determination）的說明，下列哪一項正確？

(A) 當 SSR 為 0 時，判定係數為 1.0

(B) 比較不同迴歸模型的解釋力，可直接用判定係數大

	小進行比較 (C) 判定係數代表因變數（Dependent Variable）解釋自變數（Independent Variable）的百分比 (D) 判定係數和相關係數有關係
C	50. 請問 Python 線性模型與監督式學習概念，下列哪一項正確？ (A) 線性模型只能處理二元分類問題 (B) 線性模型和非線性模型的正確率表現相同 (C) 監督式學習是指給定輸入和輸出數據集，訓練出一個能夠將輸入映射為輸出的模型 (D) 監督式學習只能用於分類問題