

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 1 頁，共 11 頁

單選題 50 題（佔 100%）

D	1. 在資料分析時常產生一些特殊值，下列何者「不是」R 語言的特殊值？ (A) NULL (B) NA (C) Inf (D) Error
A	2. 關於資料之遺缺值處理，下列敘述何者「不正確」？ (A) 無須考慮遺缺值比例，全部刪除 (B) 類別資料補上眾數之值 (C) 利用模型補上估計產生之值 (D) 透過差值法（interpolation method）補上該值
C	3. 在正規表達式中，下列何者可以完整比對出手機號碼：0912-345678？ (A) [0-9]+ (B) 09[0-9]{8} (C) [0-9]{4}-?[0-9]{3}?[0-9]{3} (D) 09{2}-[0-9]{8}
D	4. 在資料清理過程中，下列何者「不適合」用來找出極端值（outlier）或雜訊（noisy）資料？ (A) 盒鬚圖法（box plot） (B) 漢佩爾辨識法（Hampel identifier） (C) 標準化分數法（standardization） (D) 迴歸係數正規化法（regularized regression）
C	5. 關於資料前處理（data preprocessing），下列敘述何者「不正確」？ (A) 屬性尺度調整（feature scaling）可幫助大部分模型有更快的收斂與提升準確度 (B) 類別資料的處理須注意其有序與無序的特性 (C) 所有的模型建置前，事前都需要進行資料特徵縮放 (D) 類別無序資料之數值空間轉換，可用單熱編碼（one-hot encoding）方法
A	6. 下列何者在繪製時需要使用到資料的四分位數？ (A) 盒鬚圖（box plot） (B) 目標投影追蹤（targeted projection pursuit） (C) 散點圖（scatter plot） (D) 平行座標圖（parallel coordinates）
D	7. 關於資料敘述與摘要統計之內容，下列敘述何者「不正確」？

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 2 頁，共 11 頁

	<p>(A) 資料抽樣常見的有：簡單隨機抽樣、系統抽樣、分層隨機抽樣</p> <p>(B) 進行資料計算與圖表製作，例如：次數分配表、直方圖</p> <p>(C) 衡量資料集中趨勢的統計量，例如：平均數、中位數、眾數</p> <p>(D) 比較兩筆資料的分散程度，例如：相關係數</p>
C	<p>8. 在一份 100 位員工的薪資報告中，最低薪的員工薪水為 28,000 元，最高薪員工的薪水為 98,000 元，如果我們要將最低薪資的員工歸為薪資介於 20,000 ~ 29,999 元的區間，假設各區間之寬度相等，那所有員工的薪水應該分為幾個級距？</p> <p>(A) 6</p> <p>(B) 7</p> <p>(C) 8</p> <p>(D) 9</p>
D	<p>9. 關於變異係數 (Coefficient of Variation, CV) 與標準分數 (Z-Score)，下列敘述何者較「不正確」？</p> <p>(A) 變異係數 (Coefficient of Variation, CV) 為無單位數值，所以適合對兩組不同單位資料的分散程度進行比較</p> <p>(B) 標準分數 (Z-Score) 就是要把原來是不同評分尺度的分數，轉換成具有同一評分尺度的分數，以利彼此間的比較和運算之用</p> <p>(C) 變異係數 (Coefficient of Variation, CV) 適合對兩組單位相同但平均數 (Mean) 相差很大的資料進行比較</p> <p>(D) 標準分數 (Z-Score) 是將原始資料映射到[0,1]區間</p>
C	<p>10. 若要觀察不同區間 (每 50 毫米為一級距) 的降雨量頻率分佈狀態，最適合用下列何種圖表來進行呈現？</p> <p>(A) 散佈圖 (Scatter plot)</p> <p>(B) 長條圖 (Bar plot)</p> <p>(C) 直方圖 (Histogram)</p> <p>(D) 圓餅圖 (Pie Chart)</p>
A	<p>11. 下列何種分析方法需要類別標籤 (label) 資訊？</p> <p>(A) 線性判別分析 (linear discriminant analysis)</p> <p>(B) 主成分分析 (principle component analysis)</p> <p>(C) 潛在語意分析 (latent semantic analysis)</p> <p>(D) 獨立成分分析 (independent component analysis)</p>
B	<p>12. 對於某些資料屬性內出現異常大的值，有可能會出現誤導模型訓練的結果，此時會對該屬性值進行下列何種處理方式，使所有屬性值被轉換到 0 至 1 之間？</p>

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 3 頁，共 11 頁

	<p>(A) 資料組織</p> <p>(B) 資料屬性尺度調整</p> <p>(C) 資料清理</p> <p>(D) 資料分析</p>
C	<p>13. 下列何者較「不適合」用來作為屬性萃取 (feature extraction) 的方法？</p> <p>(A) 主成分分析 (principal component analysis)</p> <p>(B) 拉普拉斯特徵映射法 (Laplacian eigenmaps)</p> <p>(C) 交叉驗證 (cross validation)</p> <p>(D) 自組織映射圖 (self-organizing map)</p>
D	<p>14. 深度學習 (deep learning) 中，會透過自編碼器 (AutoEncoder, AE) 來對影像資料降維。關於自編碼器，下列敘述何者「不正確」？</p> <p>(A) 去除影像中的雜訊</p> <p>(B) 更加準確的進行影像分類</p> <p>(C) 可用來產生新的影像結構</p> <p>(D) 為一種監督式學習 (supervised learning)</p>
A	<p>15. 在文字探勘技術中，為了使文字資料轉換為電腦看得懂的數值資料，資料科學家建立一個二維結構，其中屬性為字典中所有字詞，而屬性的數量為字典中的詞彙數量。每一篇文章經過斷詞之後，會在此結構中建立一筆紀錄，判斷每一個字詞在各文章中是否出現。請問此技術運用了下列何種屬性萃取 (feature extraction) 的方法？</p> <p>(A) 單熱編碼 (one-hot encoding)</p> <p>(B) 分級裝箱 (binning)</p> <p>(C) 四捨五入 (rounding)</p> <p>(D) 對數轉換 (log transformation)</p>
C	<p>16. 關於巨量資料處理，下列敘述何者正確？</p> <p>(A) 任何資料皆可作為特定分析目的訓練樣本</p> <p>(B) 巨量資料缺少某些變量不會影響判斷結果</p> <p>(C) 可透過網路爬蟲或 API 來搜集大量外部資料</p> <p>(D) 透過巨量資料處理可由機器完全取代人工判斷</p>
D	<p>17. 關於 MapReduce 框架，下列敘述何者「不正確」？</p> <p>(A) Mapper 的輸出需要是鍵值組 (key-value pair) 的結構</p> <p>(B) 實現 Reducer，通常是定義如何處理個別鍵值下的值集合</p> <p>(C) Reducer 的輸出值通常也是鍵值組 (key-value pair) 的結構</p> <p>(D) 資料在進入 Map 階段之前會經過整理階段 (shuffle)</p>
D	<p>18. 為了能夠有效分散處理巨量資料，分散式資料處理演算法常基於下列</p>

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 4 頁，共 11 頁

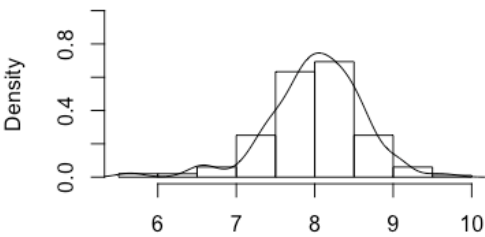
	何種概念進行設計？ (A) 貪婪演算法 (greedy) (B) 啟發法 (heuristic) (C) 反覆迭代 (iteration) (D) 先分散 (map) 後聚合 (reduce)
D	19. 假設某一企業使用大數據進行分析，下列敘述何者「不正確」？ (A) 大數據透過內外部的資料做結合 (B) 可透過大數據分析來輔助決策行為 (C) 大數據的資料結構並非固定型態 (D) 影音類型的資料無法進行應用
B	20. 關於 HDFS 之 Erasure Coding (EC) 技術，下列敘述何者「不正確」？ (A) EC 適用於節省 HDFS 總空間 (B) EC 適用於常用之資料，主要目的是提升查詢效率 (C) 當資料發生損壞，則可透過 Parity-Cell 等解碼計算，並重新恢復資料 (D) HDFS EC 的架構設計包含了 ECManager 與 ECWorker，屬於主從式架構
B	21. 參考下方報表之結果，下列敘述何者正確？ <pre>> summary(faithful) eruptions waiting Min. :1.600 Min. :43.0 1st Qu.:2.163 1st Qu.:58.0 Median :4.000 Median :76.0 Mean :3.488 Mean :70.9 3rd Qu.:4.454 3rd Qu.:82.0 Max. :5.100 Max. :96.0</pre> (A) eruptions 變數的最小值為 43.0 (B) eruptions 變數的 75 百分位數為 4.454 (C) waiting 變數的最大值為 82.0 (D) waiting 變數的中位數為 70.9
A	22. 若要將 3 種不同飲料與 4 種不同商店進行二因子變異數分析，請問交互作用的自由度為下列何者？ (A) 6 (B) 7 (C) 12 (D) 20
C	23. 下列何種統計量「不能」由盒鬚圖 (box plot) 得知？

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 5 頁，共 11 頁

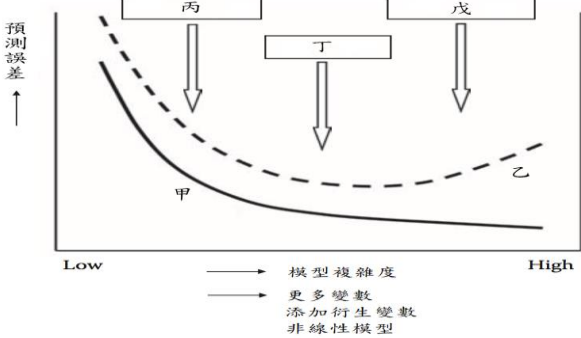
	<p>(A) 最小值</p> <p>(B) 中位數</p> <p>(C) 變異數</p> <p>(D) 全距</p>
D	<p>24. 對自變數 X 與依變數 Y 作簡單線性迴歸得到的相關係數 r，下列敘述何者正確？</p> <p>(A) $r = -1$ 代表 X 與 Y 完全無關</p> <p>(B) $r = 0$ 代表數據點恰好落在同一條水平直線上</p> <p>(C) $r > 0$ 代表 X、Y 間有因果關係</p> <p>(D) $r = 1$ 代表 $Y = aX + b$ (a、b 是常數，$a > 0$)</p>
D	<p>25. 在假設檢定中，若發生型 I 誤差之機率為 α 與發生型 II 誤差的機率為 β，下列敘述何者「不正確」？</p> <p>(A) 顯著水準為 α 之極大值</p> <p>(B) 犯型 I 誤差的嚴重性甚於犯型 II 誤差</p> <p>(C) 在同一檢定下，若 β 減少，則 α 將變大</p> <p>(D) β 越大表示檢定結果越好</p>
A	<p>26. 下圖為海藻資料集的變數 mx pH 分佈狀況，請問此圖最接近下列何種機率分佈？</p> <p>Histogram of mx pH value</p>  <p>(A) 常態分佈</p> <p>(B) 偏態分佈</p> <p>(C) 幾何分佈</p> <p>(D) 均勻分佈</p>

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期： 108 年 12 月 7 日

第 6 頁，共 11 頁

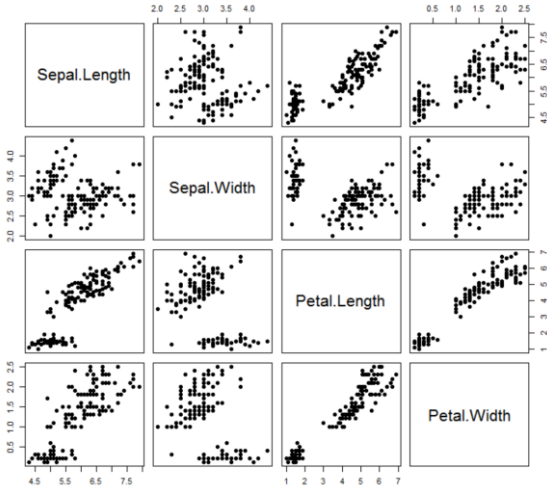
D	<p>27. 參照下圖中模型複雜度（model complexity）與預測誤差（prediction error）之間的變化關係，下列敘述何者正確？</p>  <p>(A) 實曲線甲為測試集（test set）樣本下的模型複雜度與預測誤差之間的變化關係</p> <p>(B) 虛曲線乙為訓練集（training set）樣本下的模型複雜度與預測誤差之間的變化關係</p> <p>(C) 丙段表過度配適（overfitting），戊段表配適不足（underfitting）</p> <p>(D) 丁段為較佳的模型複雜度</p>
D	<p>28. 某預測變數有兩個獨一無二的值，假設有 1000 個樣本，其中 999 個樣本的預測變數值相同，下列敘述何者「不正確」？</p> <p>(A) 類別型預測變數與數值型預測變數的退化分佈辨識方法不盡相同</p> <p>(B) freqRatio 的值為 999，它是最常見的類別值頻次，除以次常見類別值頻次的比值</p> <p>(C) percentUnique 的值為 0.002，它是以獨一無二的類別值數量與樣本大小的比值</p> <p>(D) 此變數不屬於近乎零變異（near-zero variance）的狀況</p>
C	<p>29. 假設隨機變數 X_1、X_2 獨立，且遵從相同的常態分佈 $N(\mu, \sigma^2)$。若 $Y = (X_1 + X_2)/2$，請問 Y 遵從的分佈為下列何者？</p> <p>(A) $N(\mu, 2\sigma^2)$</p> <p>(B) $N(2\mu, \sigma^2)$</p> <p>(C) $N(\mu, \sigma^2/2)$</p> <p>(D) $N(\mu/2, \sigma^2)$</p>
A	<p>30. 「薪資」資料集中的觀察值（單位千元），依遞增順序顯示為：28, 30, 32, 35, 35, 40, 45, 47, 47, 80, 90。請問上述資料中的觀察值，何者眾數值（mode）大於中位數（median）？</p> <p>(A) 47</p> <p>(B) 90</p>

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 7 頁，共 11 頁

	<p>(C) 35</p> <p>(D) 無</p>
C	<p>31. 已知 iris 資料集前 3 筆資料如下圖所示：</p> <pre>> head(iris, n=3) Sepal.Length Sepal.Width Petal.Length Petal.Width Species 1 5.1 3.5 1.4 0.2 setosa 2 4.9 3.0 1.4 0.2 setosa 3 4.7 3.2 1.3 0.2 setosa</pre> <p>請問下列的 R 語言指令選項中，何者可完成以下散佈圖矩陣？</p>  <p>(A) pairs(iris)</p> <p>(B) pairs[iris]</p> <p>(C) pairs(iris[-5])</p> <p>(D) pairs[iris(-5)]</p>
A	<p>32. 關於關聯型態探勘的特點，下列敘述何者「不正確」？</p> <p>(A) 關聯型態探勘所得到的結果，因為可以直接進行應用，所以廣受歡迎</p> <p>(B) 關聯型態分析容易從隨機的型態中妄下虛假的結論</p> <p>(C) 關聯型態探勘符合資料探勘挖掘資料庫中無預期知識的理念</p> <p>(D) 關聯型態探勘的分析方法對於小資料集的用處不大</p>
D	<p>33. 下列何種方法通常應用在集群（clustering）問題？</p> <p>(A) 支援向量機（support vector machine）</p> <p>(B) 隨機森林（random forest）</p> <p>(C) k 近鄰法（k nearest neighbors）</p> <p>(D) k 平均數（k-means）</p>
C	<p>34. 關於 k 平均數（k-means）與噪訊偵測之空間密度集群算法（Density-Based Spatial Clustering of Applications with Noise,</p>

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 8 頁，共 11 頁

	<p>DBSCAN)，下列敘述何者「不正確」？</p> <p>(A) 兩者都是集群分析</p> <p>(B) k-means 基於距離的概念，而 DBSCAN 基於密度的概念</p> <p>(C) 兩者都需要事先告知分群的數量</p> <p>(D) k-means 集群結果易受離群值的影響</p>
C	<p>35. 關於傳統「階層式集群法 (hierarchical cluster analysis)」，下列敘述何者「不正確」？</p> <p>(A) 常利用聚合法 (agglomerative approach) 和分裂法 (divisive approach) 產生所需的階層結構</p> <p>(B) 將彼此相似度高的較小群集合併成較大的群集，或者將較大群集進行分離</p> <p>(C) 利用資料點間密度的關係來分群</p> <p>(D) 利用樹狀結構圖表示群集彼此關係之分群法</p>
D	<p>36. 關於「關聯規則 (association rule)」，下列敘述何者「不正確」？</p> <p>(A) 關聯規則可以從商品交易中找到隱含的購買規則</p> <p>(B) 支持度 (support) 是衡量前提項目 (antecedent item) 與結果項目 (consequent item) 一起出現的機率</p> <p>(C) 信賴度 (confidence) 是衡量前提項目發生情況下，結果項目發生的條件機率</p> <p>(D) 增益率 (lift) 是衡量信賴度與前提項目單獨發生時二者機率比值</p>
D	<p>37. 關於 k 平均數 (k-means) 集群分析，下列敘述何者正確？</p> <p>(A) 適合解決非球形或數據密度變化大的集群問題</p> <p>(B) 演算法只要收斂，保證可以獲得最佳的集群結果</p> <p>(C) 事前不需要估算資料中有多少集群存在，即能執行算法</p> <p>(D) 不如其它集群演算法精細縝密，但在許多真實的情境下，能將集群的任務處理得足夠好</p>
A	<p>38. 屬性轉換 (feature transformation) 與資料縮減 (data reduction) 屬於資料前處理 (data preprocessing) 的重要工作，下列敘述何者正確？</p> <p>(A) 樹狀模型、最小絕對值縮減和選擇算子 (Least Absolute Shrinkage and Selection Operator, LASSO)、多變量適應性雲形迴歸 (Multivariate Adaptive Regression Splines, MARS) 等算法內嵌有變數選擇機制的方法，對於預測變數中的雜訊，或是無訊息力的變數等較不敏感</p> <p>(B) 偏最小平方法 (Partial Least Squares, PLS) 是非監督式的屬性萃取 (feature extraction)</p>

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 9 頁，共 11 頁

	(C) 最有效的變數編碼取決於數學技巧，無關於領域知識 (D) 資料前處理的需求都一樣，與後續建模所選用的模型種類無關
C	39. 屬性萃取 (feature extraction) 是指將原始資料的屬性進行結合，以產生新的代理變數 (surrogate variables)，下列常用的降維 (Dimension Reduction) 方法何者「不屬於」屬性萃取的方式？ (A) 非負矩陣分解 (non-negative matrix factorization) (B) 因子分析 (factor analysis) (C) 集群 (clustering) (D) 類神經網絡之自動編碼器 (auto-encoders)
A	40. 關於非監督式學習 (unsupervised learning)，下列敘述何者「不正確」？ (A) 線性可加模型 (General Additive Models, GAM)、效能提升模型 (Boosting) 與支援向量機 (Support Vector Machine, SVM) 等屬於非監督式學習的範疇 (B) 研究是否預測變數 (predictors) X_1, X_2, \dots, X_p 之間存在有趣的型態 (C) 研究是否能以具訊息力的方式視覺化資料背後的結構與關係 (D) 能發現變數間或觀測值間的子群體
C	41. 梯度陡降法 (gradient descent) 是機器學習中常使用的參數估計方法，可透過修正步距 (step size) α 來調整整體收斂的速度，請問若 α 過大時，會導致下列何種狀況發生？ (A) 太快收斂 (B) 收斂速度過慢 (C) 無法收斂 (D) 以上皆有可能發生
A	42. 下列學習方法，何者「難以」獲得人類容易理解的知識或特徵？ (A) 多層感知機 (multilayer perceptron) (B) 決策樹 (decision tree) (C) 羅吉斯迴歸 (logistic regression) (D) 關聯規則探勘 (association rule mining)
C	43. 關於配適不足 (underfitting)，下列敘述何者正確？ (A) 訓練誤差較大，測試誤差較小 (B) 訓練誤差較小，測試誤差較大 (C) 訓練誤差較大，測試誤差較大 (D) 訓練誤差較小，測試誤差較小
C	44. 關於下方接收者操作特性曲線 (Receiver Operating Characteristic Curve,

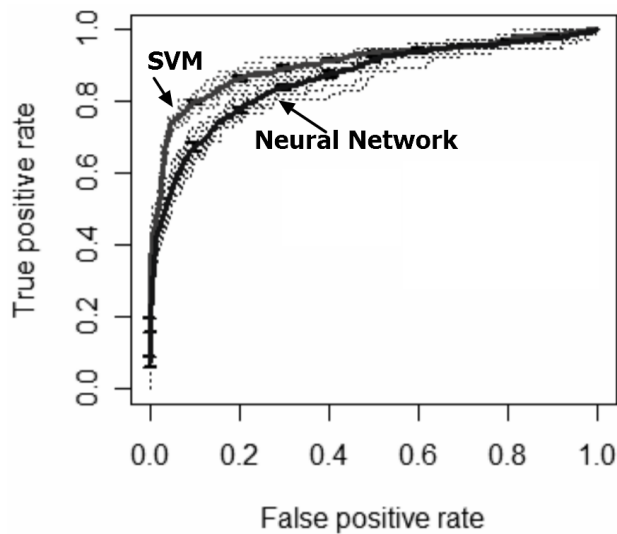
108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 10 頁，共 11 頁

ROC) 圖，下列敘述何者正確？



- (A) 假陽率 (false positive rate) 數值愈大表示分類較準確
- (B) 真陽率 (true positive rate) 數值愈大表示分類較不準確
- (C) 支援向量機 (Support Vector Machine, SVM) 模型分類準確率較類神經網路 (Neural Network, NN) 模型為佳
- (D) 上述接收者操作特性曲線無法判斷支援向量機 (Support Vector Machine, SVM) 模型或類神經網路 (Neural Network, NN) 模型的分類準確率

A 45. 關於迴歸模型，下列敘述何者「不正確」？

- (A) 可用來解釋資料現象間的因果關係
- (B) 利用自變數來預測依變數未來可能產生之值
- (C) 視其函數之型態分為線性與非線性
- (D) 根據自變數個數可分為簡單迴歸分析 (simple regression analysis) 及複迴歸分析 (multiple regression analysis)

A 46. 下方表格是針對同一份資料建立的四種複迴歸模型，根據各種模型之指標資訊，請問下列何者為最佳模型？

模型編號	模型	AIC	BIC	C_p	R^2
模型 1	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$	-55	50	3	0.8
模型 2	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$	-55	50	4	0.8
模型 3	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$	-30	60	3	0.8
模型 4	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$	10	100	3	0.8

AIC 為赤池信息量準則 (Akaike Information Criterion)；

BIC 為貝葉斯信息準則 (Bayesian Information Criterion)；

C_p 為馬洛斯 C_p (Mallows' C_p)；

108 年度初級巨量資料分析師能力鑑定試題

科目 2：資料處理與分析概論

考試日期：108 年 12 月 7 日

第 11 頁，共 11 頁

	<p>R^2 為判定係數 (coefficient of determination)</p> <p>(A) 模型 1</p> <p>(B) 模型 2</p> <p>(C) 模型 3</p> <p>(D) 模型 4</p>													
C	<p>47. 關於決策樹 (decision tree) 演算法，下列何者不侷限使用於離散型資料？</p> <p>(A) ID3</p> <p>(B) CHAID</p> <p>(C) CART</p> <p>(D) C4.5</p>													
A	<p>48. 下列何者為非監督式學習 (unsupervised learning) 演算法？</p> <p>(A) 關聯規則學習 (association rule learning)</p> <p>(B) 決策樹 (decision tree)</p> <p>(C) 天真貝氏法 (Naïve Bayes)</p> <p>(D) 隨機森林 (random forest)</p>													
B	<p>49. 關於係數正規化 (regularization) 在機器學習和深度學習中的作用，下列敘述何者「不正確」？</p> <p>(A) 防止過度配適</p> <p>(B) 降低雜訊樣本對模型的影響</p> <p>(C) 降低模型複雜度</p> <p>(D) 改變資料分佈，讓通過模型得到的資料分布和真實的資料生成過程相匹配</p>													
D	<p>50. 關於二元分類 (binary classification)，若一分類模型產生之混淆矩陣 (confusion matrix) 如下，該模型之精確度 (precision) 為下列何者？</p> <table border="1"><tr><td colspan="2" rowspan="2"></td><td colspan="2">正確答案</td></tr><tr><td>True</td><td>False</td></tr><tr><td rowspan="2">預測結果</td><td>True</td><td>8</td><td>3</td></tr><tr><td>False</td><td>12</td><td>11</td></tr></table> <p>(A) 3/11</p> <p>(B) 8/20</p> <p>(C) 19/34</p> <p>(D) 8/11</p>			正確答案		True	False	預測結果	True	8	3	False	12	11
				正確答案										
		True	False											
預測結果	True	8	3											
	False	12	11											