中央存款保險公司 108 年進用正式職員甄試試題

甄試類別【代碼】：數據分析人員-資訊類(九職等)【P2401】

專業科目 1：數據分析與演算法運用(如 R、Python)　　**＊入場通知書編號：＿＿＿＿＿＿**

注意：①作答前先檢查答案卷，測驗入場通知書編號、座位標籤、應試科目等是否相符，如有不同應立即請監試人員處理。使用非本人答案卷作答者，不予計分。
②本試卷為一張雙面，非選擇題共 4 大題，每題各 25 分，共 100 分。
③非選擇題限以藍、黑色鋼筆或原子筆於答案卷上採**橫式**作答，並請**依標題指示之題號於各題指定作答區內作答**。
④**請勿於答案卷上書寫姓名、入場通知書編號或與答案無關之任何文字或符號。**
⑤本項測驗僅得使用簡易型電子計算器(不具任何財務函數、工程函數、儲存程式、文數字編輯、內建程式、外接插卡、攝(錄)影音、資料傳輸、通訊或類似功能)，且不得發出聲響。應考人如有下列情事扣該節成績 10 分，如再犯者該節不予計分。1.電子計算器發出聲響，經制止仍執意續犯者。2.將不符規定之電子計算器置於桌面或使用，經制止仍執意續犯者。
⑥**答案卷務必繳回，違反者該節成績以零分計算。**

第一題：

在資料收集過程中往往因為各種原因導致遺漏值的情況發生，使得遺漏值處理在數據分析領域中占有極為重要之地位，請回答下列問題：

（一）何謂遺漏值？在一般情況下，處理遺漏值的作法有哪兩種？此兩種作法的選用時機為何？【5 分】

（二）下表為儲存在 test.csv 檔案中的資料數據，請以 Python 或 R 語言，將表中的所有 Null 遺漏值以直接刪去法處置後將結果顯示在螢幕上。【10 分】
※非以 Python 或 R 語言作答將不予計分。

	Count_A	Count_B	Count_C	Count_D
0	0.5	0.9	0.4	Null
1	0.8	0.6	Null	Null
2	0.6	0.4	0.8	0.8
3	0.5	0.3	Null	0.4
4	0.7	Null	0.2	0.5
5	0.5	0.8	0.5	Null

（三）承上題，請以 Python 或 R 語言，將上表中的所有 Null 遺漏值以任一種填補法處置後將結果顯示在螢幕上。【10 分】
※非以 Python 或 R 語言作答將不予計分。

第二題：

關聯法則 Association Rule（俗稱購物籃分析）是實務上常用的資料探勘演算法的一種，透過法則的探勘與萃取，分析者便能夠清楚得知兩事務共同發生之潛在關聯性。例如「某銀行資料分析師自資料庫中探勘出信用卡顧客之應繳總金額大於$10,000 元且截止日前提早繳款次數大於 5 次者，較容易接受該行的電話促銷方案」。請回答下列問題：

（一）關聯法則在運作過程中高度仰賴哪兩種門檻值？請簡述此兩種門檻值的用途。【10 分】

（二）請以 Python 或 R 語言，撰寫出以關聯法則演算法為基礎之可執行程式碼。【15 分】
※答案內容需依照下圖資料庫表單為設計參照，其中 1 表示有買，0 表示沒買。

交易編號	麵包	牛奶	尿布	雞蛋	可樂
A001	1	1	1	1	0
A002	0	1	1	1	1
A003	1	1	1	0	1
A004	1	0	1	1	0
A005	1	1	0	0	0

【請接續背面】

第三題：

現有一份資料檔案如下表，包括下列欄位：身份識別(ID)：文字；性別(Gender)：男(M)/女(F)；抽菸量(SmokeAmount)：正整數或 0；睡眠時間(SleepTime)：正整數或 0；肝癌篩選(LSC)：陽性(P)/陰性(N)。此外，各個欄位資訊有可能遺失(NA)。檔案以 CSV 格式儲存(data.csv)。

ID	Gender	SmokeAmount	SleepTime	LSC
A0001	M	NA	4	P
A0002	F	0	8	N
A0003	M	NA	3	P
…	…	…	…	…

請以 R 或 Python 語言撰寫程式：

（一）讀取 data.csv 檔案。【5 分】

（二）計算可用資料列數。（亦即每一欄位均非 NA）【5 分】

（三）計算各數字型欄位(SmokeAmount 以及 SleepTime)的平均值與標準差。（計算過程中需要忽略欄位值為 NA 的資料）【7 分】

（四）計算各類別型欄位(Gender 以及 LSC)的總次數與比例。（計算過程中需要忽略欄位值為 NA 的資料）【8 分】

下列為輸出範例：

可用資料列數：500 筆

SmokeAmount(avg/std):8.5(1.32)

SleepTime(avg/std):4.1(2.01)

Gender(Total/Percentage):M:300(60%),F:200(40%)

LSC(Total/Percentage):P:100(20%),N:400(80%)

第四題：

已知單純貝氏分類器(Gaussian Naive Bayes Classifier)運用下列條件機率公式來進行資料分類：

$$p(\text{class}|\text{data}) = \frac{p(data|class) \times p(class)}{p(data)}$$

其中 class 是類別，data 是資料，並假設資料的各項欄位之間是彼此獨立的。因此，若一資料 d 算出 $p(c_1|d) > p(c_2|d)$ ，那麼便將 d 歸類為類別 c_1。（提示：依上述公式，實際上不需計算 p(data)）

現有一份資料檔案如下表，包括下列欄位：

1.身份識別(ID)：文字

2.性別(Gender)：男(M)／女(F)

3.抽菸量(SmokeAmount)：無(NONE)／少(FEW)／多(MANY)

4.睡眠時間(SleepTime)：正常(NORMAL)／過少(FEW)／過多(MANY)

5.肝癌篩選(LSC)：陽性(P)／陰性(N)

此外，假設各個欄位資訊沒有遺失值。檔案以 CSV 格式儲存(data.csv) 。

ID	Gender	SmokeAmount	SleepTime	LSC
A0001	M	NONE	FEW	P
A0002	F	FEW	MANY	N
A0003	M	MANY	NORMAL	P
…	…	…	…	…

請以 R 或 Python 語言撰寫一程式運用單純貝氏分類器，依照上述資料檔案進行學習，之後可以對新案例(Gender, SmokeAmount, SleepTime)進行預測該案例的肝癌篩選結果是陽性(P)／陰性(N) 。【25 分】