Sleep Health and Lifestyle -SAS viya

*a*資料來源

本研究使用 Kaggle 上的「Sleep Health and Lifestyle Dataset」(作者:Laksika Tharmalingam，授權:CC0 公領域)。
此資料為合成資料(synthetic)，涵蓋性別、年齡、職業、睡眠時數、睡眠品質、運動、壓力、BMI 類別、血壓、心率、日走步數與睡眠障礙(None／Insomnia／Sleep Apnea,
共 200 筆、13 欄

link : https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

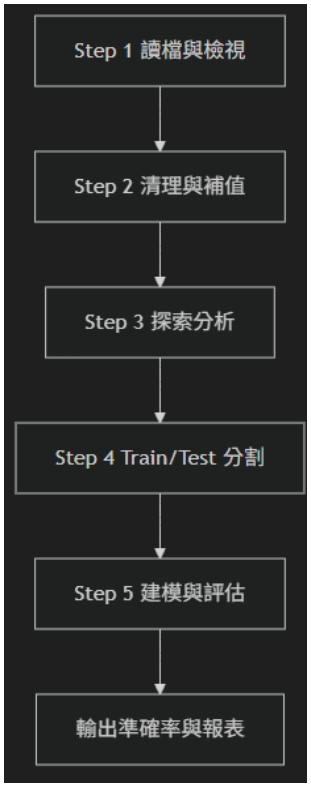
研究目的

1.建立基準分類模型，預測個體睡眠狀態(None／Insomnia／Sleep Apnea)

2.量化各特徵(睡眠時數、壓力、運動、血壓等)對預測的貢獻，找出關鍵影響因子

研究流程

流程圖：



環境設置:
平台:
SAS Viya(站台:ENGAGE PLATFORM AIOT TRIAL 2025)
介面:
SAS Studio(穩定版本 2025.08);VS Code(SAS 擴充)輔助
程式與路徑:
`/Public/Sandy/sleep_test/procdata.sasnb`(SAS Drive → Content)(程式碼詳請可見附錄1)
重現性:
隨機種子 1234;資料以 FILESRVC 讀取;不依賴 CAS 表(僅需可使用 PROC HPSPLIT)
(將「環境快照」輸出為 `env_snapshot.csv` 存回原路徑作為留證)
(說明:本專案以 Compute 工作階段執行;
如需查詢 CAS 版本/節點狀態，可先建立 CAS 連線後再執行
```bash
proc cas;
builtins.about;
quit;
```
若未啟用 CAS，報告仍可重現，因程式主要依賴 FILESRVC 與 Compute 端程序)

Step 1 讀檔與檢視：

輸入：
`/Public/Sandy/sleep_test
/sleepdata1.csv`
（SAS Content）

資料擴充（**DataMaker**）：
以「有放回抽樣」(URS) 方式將原始約 **200** 筆擴充到**2,000,000筆**，此作法保留原始類別與數值分布（屬於複製抽樣，非合成新樣本）

The SURVEYSELECT Procedure

Selection Method	Unrestricted Random Sampling

Input Data Set	RAW
Random Number Seed	1234
Sample Size	2000000
Expected Number of Hits	10000
Sampling Weight	0.0001
Output Data Set	RAW_BIG_TMP

Step 2 清理與補值：

1.去除干擾欄位：刪掉 `FOLD` 欄位(此欄位為紀錄資料合併來源檔)

2.補缺值：使用 中位數 補 `Sleep_Duration`、`Quality_of_Sleep` 的缺漏。

3.新增風險標記：依血壓產生 `BP_Abnormal`：

- 若 `Systolic ≥ 130` 或 `Diastolic ≥ 85` → `Yes`
- 否則 → `No`；若缺值 → 留空

處理前

The MEANS Procedure

Variable	N	N Miss
Person_ID	2000000	0
Age	2000000	0
Sleep_Duration	1920669	79331
Quality_of_Sleep	1969632	30368
Physical_Activity_Level	2000000	0
Stress_Level	2000000	0
Systolic	2000000	0
Diastolic	2000000	0
Heart_Rate	2000000	0
Daily_Steps	2000000	0
fold	2000000	0
RowID	2000000	0

處理後(遺失值、刪除欄位)

The MEANS Procedure

Variable	N	N Miss
Person_ID	2000000	0
Age	2000000	0
Sleep_Duration	2000000	0
Quality_of_Sleep	2000000	0
Physical_Activity_Level	2000000	0
Stress_Level	2000000	0
Systolic	2000000	0
Diastolic	2000000	0
Heart_Rate	2000000	0
Daily_Steps	2000000	0
RowID	2000000	0

Step 3 探索分析（**EDA**）：

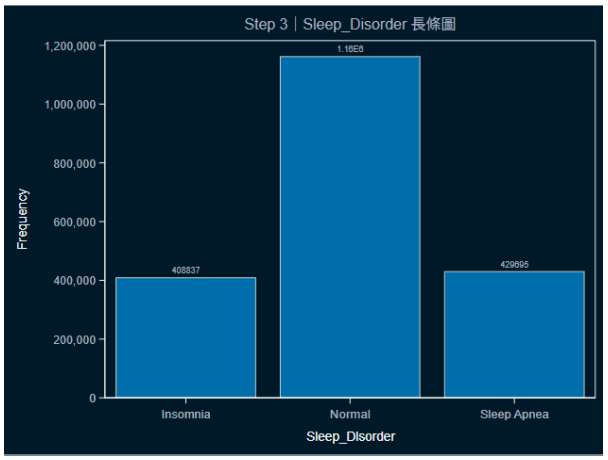
按照敘述統計方式完成筆數、平均、標準差、最小/最大、中位數、四分位數

The MEANS Procedure

Variable	N	N Miss	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Person_ID	2000000	0	195.8741275	105.2984924	1.0000000	105.0000000	201.0000000	284.0000000	374.0000000
Age	2000000	0	42.8035875	8.5002236	27.0000000	36.0000000	43.0000000	50.0000000	59.0000000
Sleep_Duration	2000000	0	7.1916734	0.7684833	5.8000000	6.5000000	7.2000000	7.8000000	8.5000000
Quality_of_Sleep	2000000	0	7.3915150	1.1477124	4.0000000	6.0000000	8.0000000	8.0000000	9.0000000
Physical_Activity_Level	2000000	0	61.3851700	20.7006346	30.0000000	45.0000000	60.0000000	75.0000000	90.0000000
Stress_Level	2000000	0	5.3183435	1.7336209	3.0000000	4.0000000	5.0000000	7.0000000	8.0000000
Systolic	2000000	0	128.8717400	7.7822840	115.0000000	125.0000000	130.0000000	135.0000000	142.0000000
Diastolic	2000000	0	85.0204245	6.2286170	75.0000000	80.0000000	85.0000000	90.0000000	95.0000000
Heart_Rate	2000000	0	69.9384680	3.7556372	65.0000000	68.0000000	70.0000000	72.0000000	85.0000000
Daily_Steps	2000000	0	6967.49	1560.49	3000.00	6000.00	7000.00	8000.00	10000.00
RowID	2000000	0	1000000.50	577350.41	1.0000000	500000.50	1000000.50	1500000.50	2000000.00

目標變數分布：看 Sleep_Disorder 各類別的件數

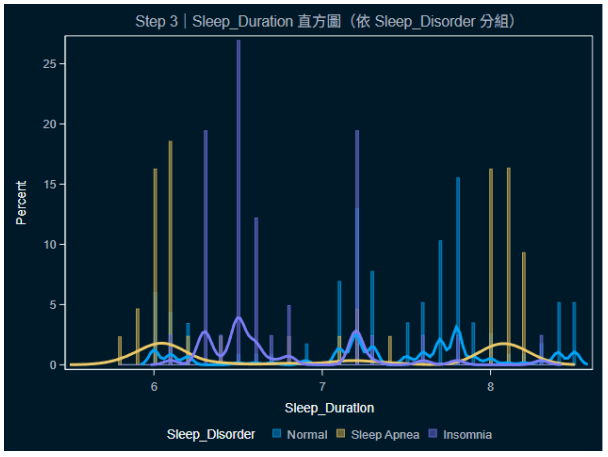
*Sleep_Disorder 長條圖：一眼看哪一類最多



Normal樣本最多，約一百多萬筆；
Insomnia和Sleep Apnea各只有四十多萬筆，約為 Normal 樣本的33%。
說明資料不平均，Normal 遠多於其他兩類

建議:
在切訓練/測試集時改用分層抽樣來固定各類別比例，避免簡單隨機抽樣造成的比例漂移，讓訓練與評估更穩定、更接近真實分佈

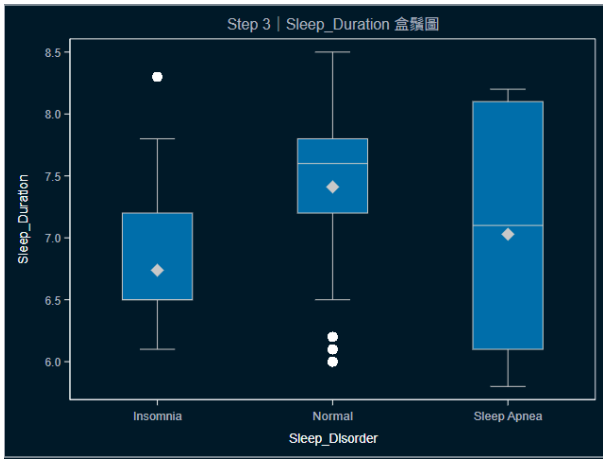
*Sleep_Duration 直方圖 + 密度（依類別分組）：觀察不同睡眠障礙類別的時數分布差異



1.Normal 的睡眠時間多集中在 7–8 小時，右邊比較高
2.Insomnia（紫色）偏短，常落在 6–7 小時，曲線低而分散
3.Sleep Apnea（卡其色）在 6–6.5 小時附近有一堆尖峰，也出現零星較長的睡眠

正常睡的人大多睡得比較久；失眠較短且不穩定；睡眠呼吸中止偏短、分布不均

*Sleep_Duration 盒鬚圖（依類別）：比較各類的中位數與離群值



正常組睡得最久、最穩，中央值大約7.5小時；
失眠組最短，落在6.5～7小時附近；睡眠呼吸中止介於兩者之間，但波動最大，同時有很多很短也有偏長的夜晚。
白色菱形是平均值、上下鬚與散點顯示離群與變異

Gender × Sleep_Disorder 列聯表：看性別與睡眠障礙的關聯概況

Step 3 | Gender × Sleep_Disorder（列聯表）

The FREQ Procedure

Frequency	Table of Gender by Sleep_Disorder			
	Sleep_Disorder			
Gender	Insomnia	Normal	Sleep Apnea	Total
Female	219770	400868	399688	1020326
Male	189067	760600	30007	979674
Total	408837	1161468	429695	2000000

男女在「是否有睡眠障礙」的分布有顯著差異（χ^2=431,123.37，df=2，p<0.001）

女生:失眠21.5%、正常39.3%、睡眠呼吸中止39.2%；
男生:失眠19.3%、正常77.6%、睡眠呼吸中止只有約3.1%

Step 4 分割：

依類別分層隨機切資料：
> 用 `Sleep_Disorder` 分層，做 `train` 80%(1,600,001筆)與 `test` 20%(399,999筆) 切分

方法：在每個類別內做簡單隨機抽樣 {‘`method’:’srs’`, ‘可重現’:{‘seed’:’1234’}}

Step 5 建模與評估：

下圖呈現 各模型的「準確率（績效）」與「運行時間」表格，在
所有達到最高準確率（**94.45%**）的模型中，**MaxDepth=40、LeafSize=150**
測試評分時間最短（138 ms），
因此在不犧牲準確度下是最合適的選擇。

決策樹—時間（毫秒）與準確率（累計結果）

Model	Prune	MaxDepth	LeafSize	TrainMS	ScoreTrainMS	ScoreTestMS	TrainAcc	TestAcc
Decision Tree	OFF	1	1	1,415	465	112	73.05%	73.03%
Decision Tree	OFF	2	2	1,709	525	143	90.50%	90.46%
Decision Tree	OFF	3	3	2,854	580	228	93.00%	92.96%
Decision Tree	OFF	4	5	2,910	596	245	93.00%	92.94%
Decision Tree	OFF	6	7	3,950	601	195	93.50%	93.44%
Decision Tree	OFF	8	10	5,228	596	201	94.50%	94.45%
Decision Tree	OFF	10	15	5,866	778	352	94.50%	94.45%
Decision Tree	OFF	12	20	5,817	646	231	94.50%	94.45%
Decision Tree	OFF	15	30	5,893	675	219	94.50%	94.45%
Decision Tree	OFF	20	50	5,927	585	140	94.50%	94.45%
Decision Tree	OFF	25	75	6,118	728	406	94.50%	94.45%
Decision Tree	OFF	30	100	5,859	692	149	94.50%	94.45%
Decision Tree	OFF	40	150	5,861	593	138	94.50%	94.45%
Decision Tree	OFF	50	200	5,854	577	173	94.50%	94.45%
Decision Tree	OFF	75	300	5,481	645	320	94.50%	94.45%
Decision Tree	OFF	100	500	2	605	147	94.50%	94.45%
Decision Tree	OFF	150	50	2	647	329	94.50%	94.45%
Decision Tree	OFF	200	5	2	599	144	94.50%	94.45%
Decision Tree	OFF	300	2	4	725	205	94.50%	94.45%
Decision Tree	OFF	500	1	2	579	139	94.50%	94.45%

混淆矩陣（實際 × 預測）及準確率

The HPSPLIT Procedure

Model-Based Confusion Matrix

Actual	Predicted			Error Rate
	Insomnia	Normal	Sleep Apnea	
Insomnia	72000	8000	2000	0.1220
Normal	2055	224141	6049	0.0349
Sleep Apnea	1904	1891	81961	0.0443

Model-Based Fit Statistics for Selected Tree

N Leaves	ASE	Mis-class	Entropy	Gini	RSS
17	0.0331	0.0547	0.2885	0.0993	39703.9

此決策樹有 17 個葉節點，整體錯誤率約 5.47%（準確率約**94.53%**）

各類別的錯誤率：Insomnia 12.2%、Normal 3.49%、Sleep Apnea 4.43%

模型整體表現符合預期**(0.945>0.6)**，主要混淆發生在「異常」兩類（失眠/呼吸中止）與 Normal 之間，特別是失眠較容易被判成正常

變數重要性**(Features)**：

Variable Importance			
	Training		
Variable	Relative	Importance	Count
BMI_Category	1.0000	315.7	1
Occupation	0.7671	242.2	2
Systolic	0.3997	126.2	2
Gender	0.2660	83.9835	3
Quality_of_Sleep	0.2007	63.3613	1
Age	0.1794	56.6297	3
Sleep_Duration	0.1313	41.4498	4

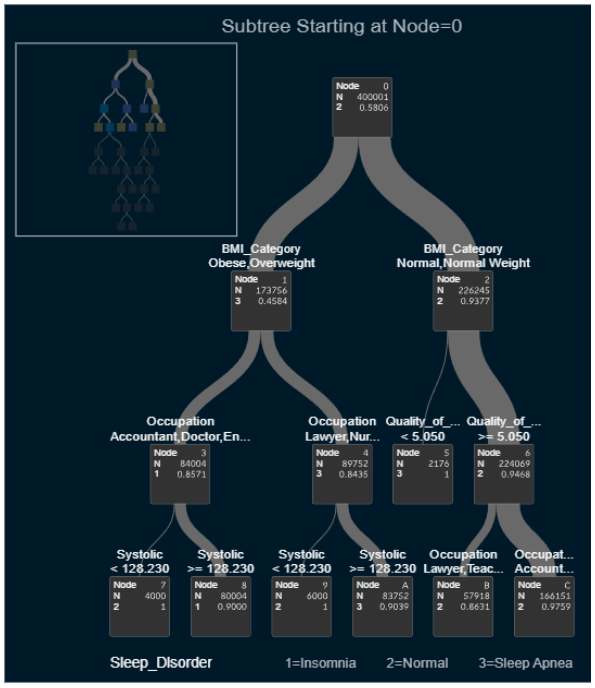
數值越大、越靠上，對模型判斷睡眠疾患越關鍵；相對重要性以 **BMI_Category=1.00** 當基準

BMI_Category（1.00）：最有影響力，是模型的第一關鍵因子

Occupation（0.77）：影響力約為 BMI 的 **77%**，排第二

Systolic（0.40）：收縮壓也很重要，約為 BMI 的 **40%**

決策樹 節點：



node1: BMI 類別，體重在「正常/偏瘦」多半被判為正常睡眠；「 過重/肥胖」一側風險明顯提高

node2:職業與睡眠品質：在正常/偏瘦的人中，只要睡眠品質≥5，就幾乎都是正常；品質差或特定職業（例如律師、護士、老師等）才會轉向失眠或睡呼吸中止

node3:血壓（Systolic）是最後的重要分叉，高於約 128 的族群更容易被判為失眠或睡呼吸中止

*底部標示 1=Insomnia、2=Normal、3=Sleep Apnea，框內的比例顯示每個葉節點的預測類別機率。