接下來為各位簡要介紹本研究的資料來源、研究目的、環境設置，以及模型流程與主要發現。

首先是資料來源。
使用 Kaggle 的 Sleep Health and Lifestyle Dataset，資料集共兩百筆、十三個欄位，涵蓋性別、年齡、職業、睡眠時數等，以及目標變數「睡眠障礙」。

研究目的有兩點。
第一，建立基準的三分類模型，預測個體的睡眠狀態：None、Insomnia、以及 Sleep Apnea。這三類分別代表，None為沒有明顯睡眠問題、Insomnia指入睡困難、Sleep Apnea，也就是睡眠呼吸中止。

第二，量化各特徵對預測結果的貢獻，從睡眠時數、壓力、運動與血壓等變數中找出關鍵影響因子。

本研究使用 SAS Viya平台，並以 VS Code 的 SAS 擴充進行輔助。

以下是研究流程與結果重點。

首先將檔案放入viya平台中，進行讀檔與檢視，並進行資料清理，利用探索分析看資料型態，並將資料切分成訓練以及測試資料集，以利最後的建模與評估。

第一步，讀檔與檢視。
先將檔案上傳至SAS Content中，並從中讀取資料。為了讓訓練與評估更穩定，我們使用 DataMaker 進行有放回抽樣，將原始兩百筆資料擴充到兩百萬筆。

第二步，清理與補值。
我們移除對模型無幫助僅記錄用的 FOLD 欄位，接著以中位數補齊 Sleep_Duration 與 Quality_of_Sleep 兩個欄位的缺漏，並依血壓閾值新增 BP_Abnormal 風險標記。

第三步，探索式分析。
根據描述性統計顯示整體樣本不平衡：Normal 最多，約一百多萬筆；Insomnia 與 Sleep Apnea 各約四十多萬筆，約為 Normal 的三分之一。
睡眠時數分布方面，Normal 多集中在七到八小時；Insomnia 偏短且分散；Sleep Apnea 介於兩者之間但波動較大。
盒鬚圖亦顯示：正常組最久、最穩；失眠組最短；睡眠呼吸中止變異最大。
在性別與睡眠障礙的列聯分析中，男女在是否有睡眠障礙的分布呈現顯著差異，提示性別與睡眠型態之間可能存在結構性的差異。

第四步，資料切分。
我們以 Sleep_Disorder 為分層變數，進行八比二的訓練與測試切分，總計約一百六十萬筆進行訓練、四十萬筆測試資料，並設定隨機種子，固定模型出來結果。

第五步，建模與評估。
基準模型採用可解釋的決策樹，並比較不同深度與葉節點大小的組合。
在所有達到最高準確率約94.45%的組合中，最大深度40、葉節點大小150，在測試階段的評分時間最短，約一百三十八毫秒，因此在不犧牲準確度的前提下為最合適的選擇。
整體錯誤率約為5.47%；各類別錯誤率依序為：Insomnia 約12.2%、Normal 約3.49%、Sleep

Apnea 約4.43%。
主要混淆發生在失眠與正常之間，顯示提升失眠類的召回率是下一步優化的重點。

在變數重要性方面，BMI_Category 的相對重要性最高；Occupation 排名第二；Systolic 收縮壓排名第三。

決策樹的規則也具可解釋性：
首先，BMI 類別是首要分叉，正常或偏瘦者多半被判為正常睡眠；過重或肥胖一側，風險明顯提高。
接著，在正常或偏瘦族群中，若睡眠品質分數大於等於五，幾乎皆為正常；若品質偏差或特定職業屬性，則更傾向失眠或睡眠呼吸中止。
最後，收縮壓高於約一二八者，更容易被判為失眠或睡眠呼吸中止。
最終樹模型包含十七個葉節點，葉節點框內的比例顯示各節點的預測機率分布。

綜合而言，我們完成了一個可重現、可解釋、且具備良好基準表現的三分類模型。
下一步將聚焦在不平衡處理、機率校準與監測指標的完善，特別是提升失眠類的辨識能力，同時維持整體準確度與臨床可解釋性。
以上是本研究的重點摘要，謝謝各位。