113 學年度

資料探勘

Term Project Report

第 5 組

313554024 林慧旻

313581011 洪明祺

313551099 李以恩

313551135 林念慈

1、 計畫目標的問題 (Target problem)

- 目標問題描述 （需包含資料輸入/處理過程/輸出）

  目標問題：我們使用了 Sleep Health and Lifestyle Dataset 作為輸入資料，試圖尋找包含與睡眠及日常生活習慣相關的多種變數，用以預測病人是否患有任何睡眠障礙。

  資料輸出：預測某人是否患有睡眠障礙，總共分為三類：無（None）、失眠（Insomnia）、睡眠呼吸中止症（Sleep Apnea）。

- 評估指標
  - Accuracy（準確率）
  - Precision（精確率）
  - Recall（召回率）
  - F1 Score

- 原預估之模型效能與目標（需對應於上節之「評估指標」）

  Baseline 模型效能：

  Accuracy: 83.19%

  Precision: 83.42%

  Recall: 83.18%

  F1 score: 83.21%

  目標：透過進階模型（如 Decision Tree、Random Forest、XGBoost 等）提升效能，模型的平均效能達到 90% 或以上。

2、 選用的資料集描述 (Descriptions of selected datasets)

- 資料集來源（需提供 Reference）
  - 來源平台：Kaggle
  - 資料集名稱：Sleep Health and Lifestyle Dataset
  - 下載連結：Kaggle 資料集

- 資料集相關描述（如：資料欄位、型態、總筆數、年份等）
  - 資料量：

  總筆數：374 筆，欄位數：13 個。

  - 欄位描述：
    - Person ID：每位個體的唯一識別碼（整數）。

- Gender：性別（字串類型，男性/女性）。
- Age：年齡（整數）。
- Occupation：職業或專業（字串類型）。
- Sleep Duration：每日睡眠時長（整數，單位為小時）。
- Quality of Sleep：睡眠質量主觀評分（整數，1-10）。
- Physical Activity Level：每日體力活動時間（整數，單位為分鐘）。
- Stress Level：壓力等級（整數，1-10）。
- BMI Category：BMI 分類（字串類型，包括 Underweight, Normal, Overweight）。
- Blood Pressure：血壓（整數，收縮壓/舒張壓）。
- Heart Rate：心率（整數，每分鐘心跳數）。
- Daily Steps：每日行走步數（整數）。
- Sleep Disorder：睡眠障礙分類（字串類型，None, Insomnia, Sleep Apnea）。

○ 資料年份：Kaggle 來源網站上並沒有提供關於更詳細的病人背景，蒐集場所，以及蒐集數據的年份。

3、 針對問題設計的分析流程（Analysis workflow）

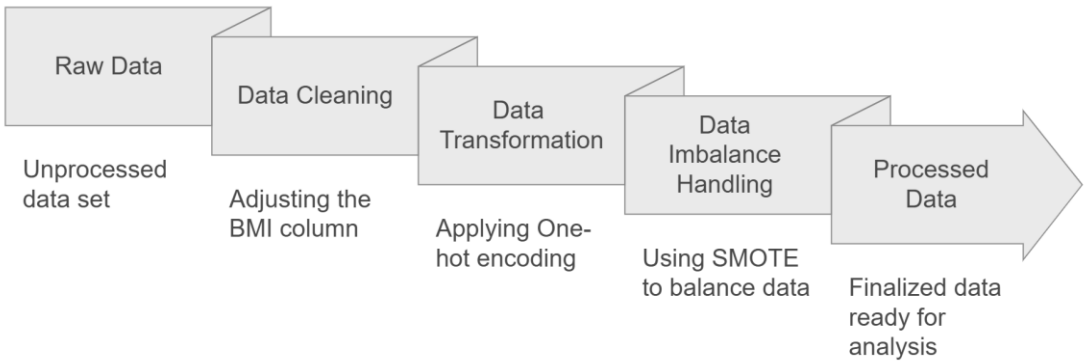

Figure 1. Analysis workflow

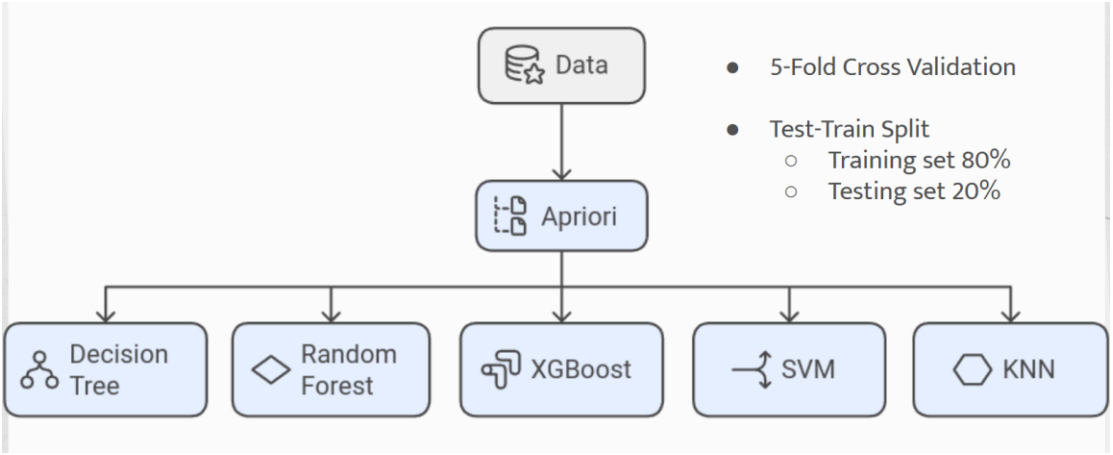

Figure 2. Training model

- 選用之資料探勘方法以及選用原因

  因為之前 lab 1 為頻繁集查找，那時選用 mlxtend 的 FP growth，但發現這個 library 的 apriori 比 FP growth 還快，所以這次選擇使用 apriori。除了效率以外，因 apriori 為比較基礎簡單的做法有許多 library 可以選擇使用、而資料集中有部分離散數(非離散以 range 拆分，就算連續資料沒有到 range 覆蓋數字過長的問題，因最少連續資料有 6 種，所以拆分成六個 range，如下圖所示)此特性適合使用 apriori。

```
Age: 30 unique values
Sleep Duration: 27 unique values
Quality of Sleep: 6 unique values
Physical Activity Level: 16 unique values
Stress Level: 6 unique values
Heart Rate: 18 unique values
Daily Steps: 19 unique values
Blood Pressure systolic: 18 unique values
Blood Pressure diastolic: 16 unique values
```

Figure 3. Number of unique values of different data

- 我們首先使用長條圖表示各個資料的分布狀態，且區分為數值型資料及類別型資料，如下圖所示：

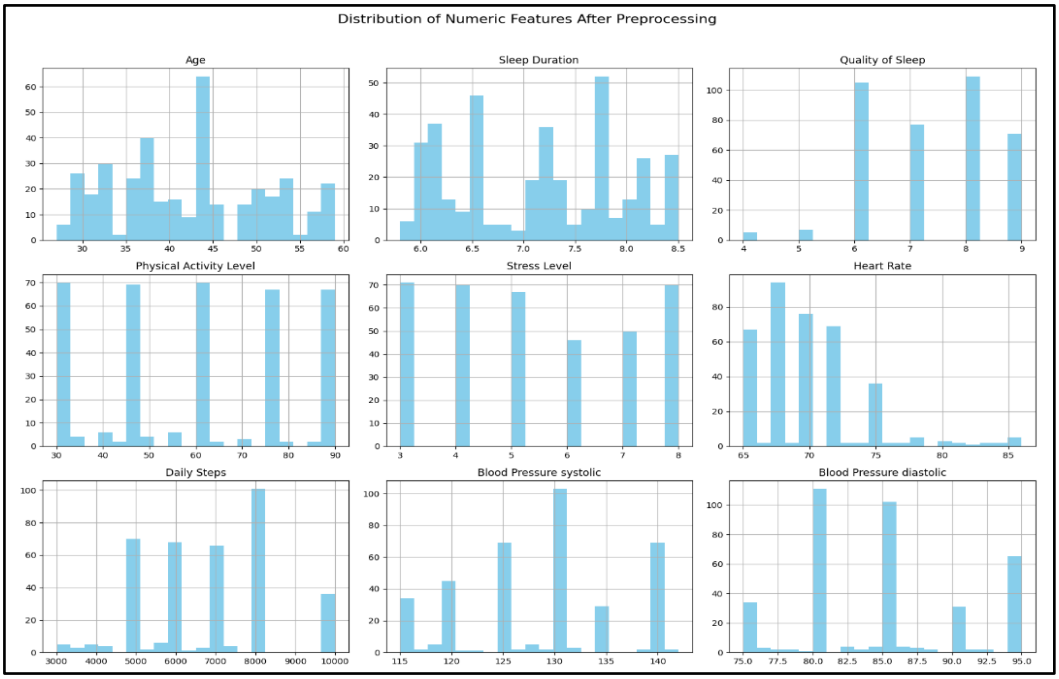

Figure 4. Distribution of numeric features after preprocessing

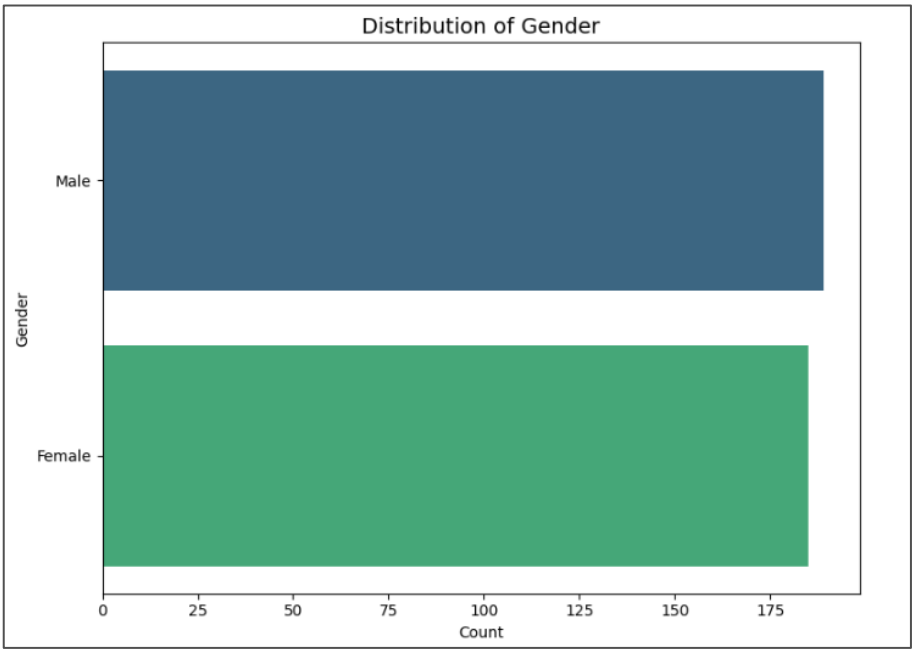

Figure 5. Distribution of gender

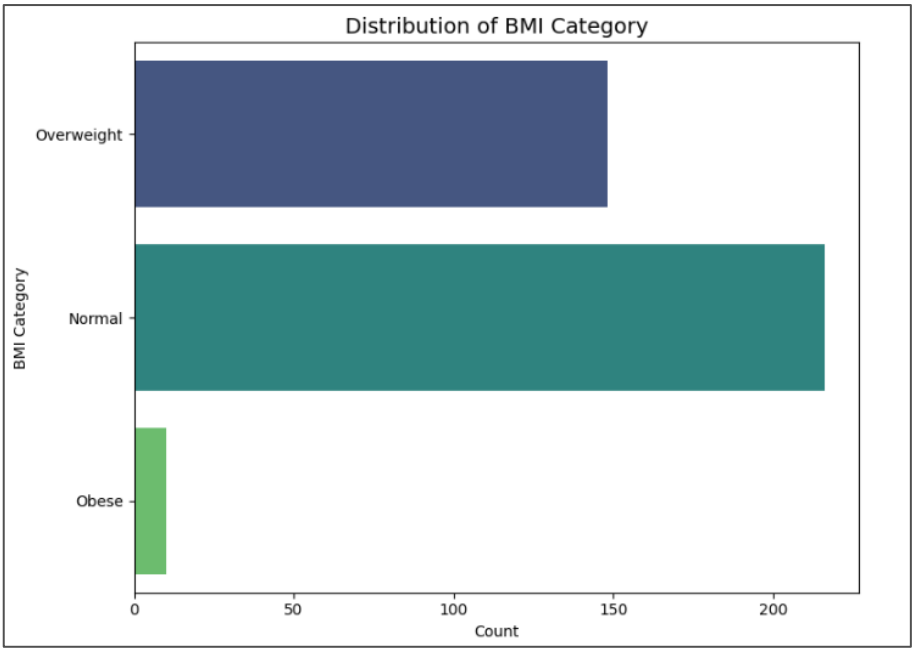

Figure 6. Distribution of BMI category

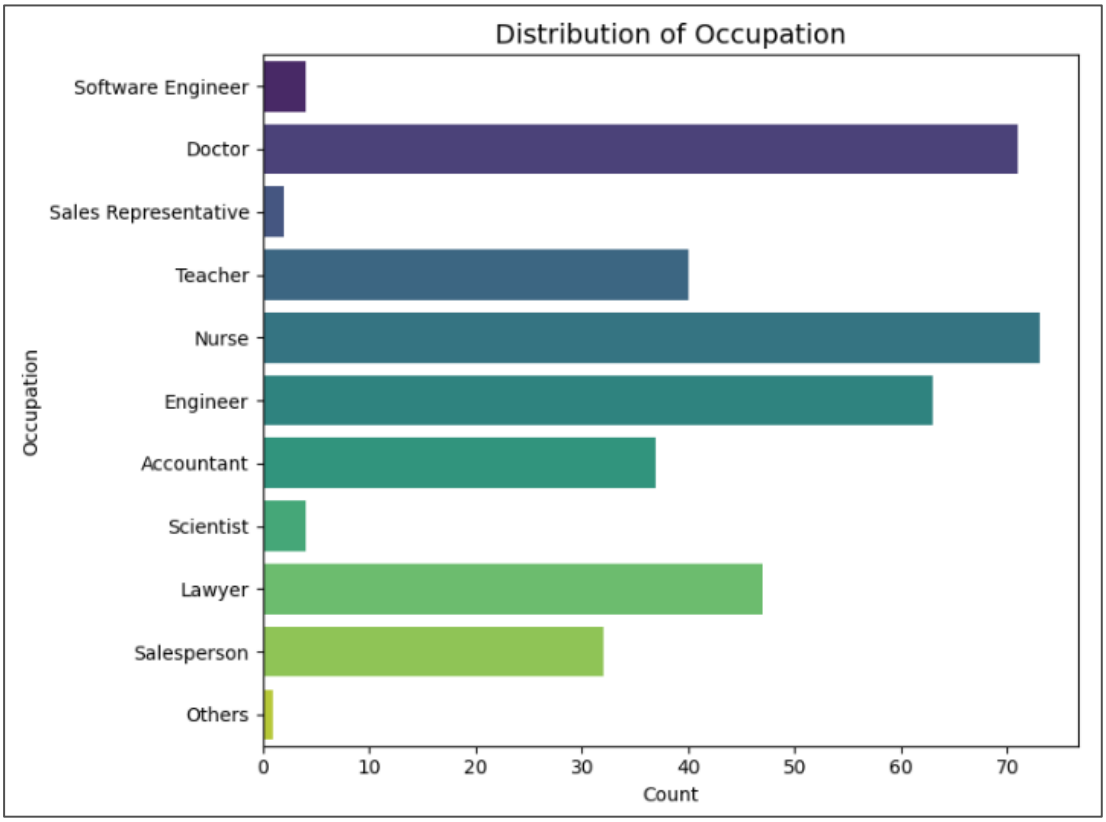

Figure 7. Distribution of occupation

在這些長條圖中，我們可以發現資料集中性別平衡、BMI 過輕的人佔少數、資料集前三大職業為醫生、護理師及工程師。另外，年齡分布在四十到五十歲居多、睡眠品質以六到九分居多，而在心率、血壓、一天的步數中，多數人皆落在正常範圍。

- 在資料前處理的步驟中，我們去除 personID，修改 BMI 中的類別（把”Normal Weight” 和 “Normal” 統一為”Normal”）。之後，我們針對資料作轉換，對於數值型資料我們做標準化的處理；對於類別型的資料我們做 One-hot encoding 以利後面的模型訓練。我們也發現資料類別有不平衡的狀況（None : Sleep Apnea : Insomnia = 6 : 2 : 2），因此我們採用 SMOTE 進行 oversampling。

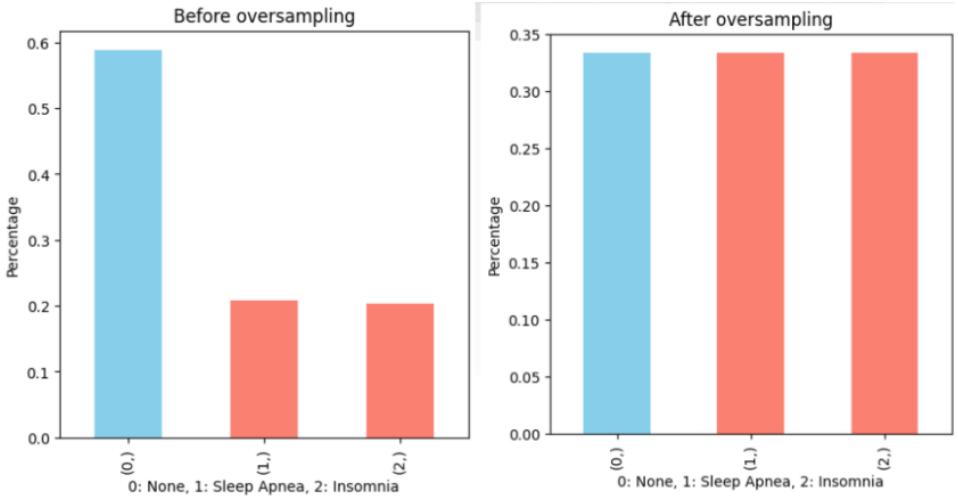

Figure 8. The process of oversampling

- 接著，我們進行 Apriori 演算法來找出頻繁子集及關聯規則，以了解睡眠障礙與什麼樣的因素有高度關聯性，在支持度最高的組合中，我們發現 BMI 正常的人通常都不會有睡眠障礙的問題、女性通常有 Sleep Apnea 的問題、BMI 超重的人會有睡眠呼吸中止症的問題等等，如下圖所示：

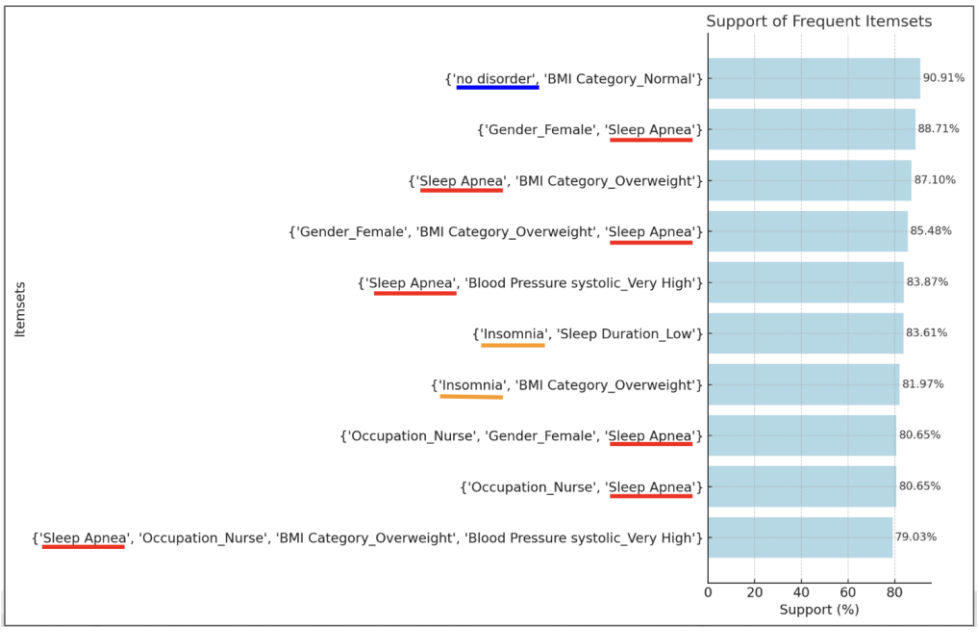

Figure 9. Apriori algorithm

之所以使用 Apriori 演算法來分析的原因是因為可以透過分析來找出睡眠習慣、睡眠習慣、環境因素和健康狀態之間的重要關聯，有了這樣的分析結果，可以幫助醫療方面的專家去理解病人有睡眠障礙的因素，進而提供病患個人化的建議；除此之外，一般人也能透過這樣的分析結果去思考自己的生活方式、環境、作息是否為造成睡眠障礙的主因，不只提供解決睡眠障礙的決策也能夠預防自己落入睡眠障礙的無底洞。

另外，我們也分析 Feature importance，如下圖為 Random Forest 所分析的特徵重要圖：

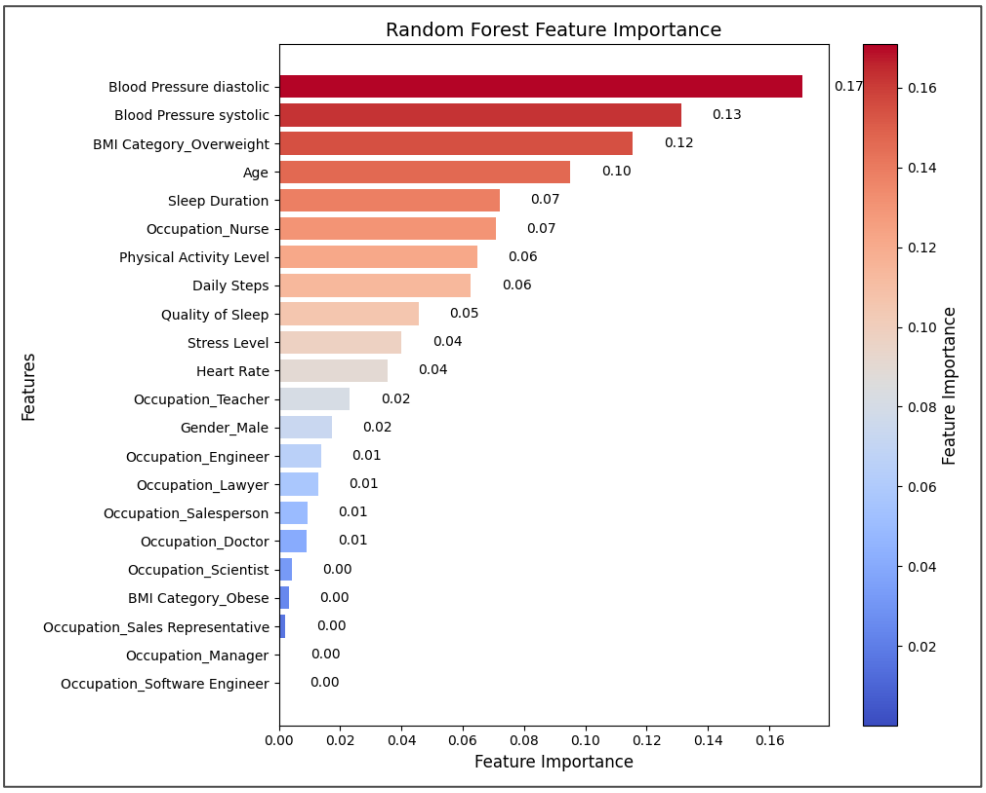

Figure 10. Feature importance obtained by RF

我們可以發現舒張壓、收縮壓、BMI 為這個模型中較重要的特徵，而每個模型的特徵重要程度可能會不一樣，因為各個模型結構不一樣、對於重要特徵的定義及標準也不一樣，因此會產出不同的特徵重要性之結果，如下圖為 Desicision Tree 的重要特徵圖，我們可以發現其相異性。

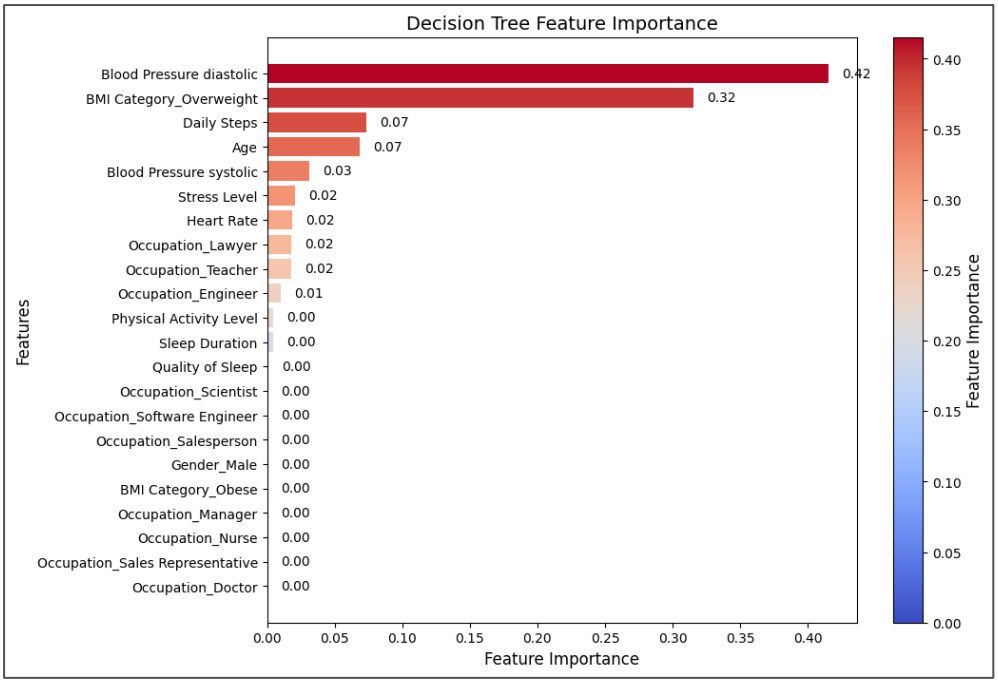

Figure 11. Feature importance obtained by decision tree

在這個重要特徵圖中則顯示，舒張壓和 BMI 過重的人是重要特徵，且很多特徵重要性的數值為 0，原因是因為這些特徵可能沒有用於 decision tree 中分裂的節點，相較於非零的特徵，這些特徵都沒有被用到，因此忽略了這些特徵。也就是說，在 Decision Tree 中只需要注意分裂的節點採用的特徵是哪些即可，這些節點才是最後預測測試集所需要被考慮的因素，這些特徵相較於特徵重要度為零的特徵，具有較大的貢獻度。

- 接著我們實作模型的訓練，因為想比較不同模型所呈現出的結果差異，所以我們採用五個模型，分別為：Decision Tree、Random Forest、XGBoost、SVM 和 KNN。在模型訓練中，我們隨機切割資料，讓訓練集：測試集為 8：2。在過程中，我們也利用 K-Fold Cross Validation 以驗證模型的可靠性，且我們的資料集資料數較少，利用 K-Fold Cross Validation 能夠確保每筆資料都被訓練及驗證過，以提高每筆資料的使用率。在每個模型中，我們分析四個結果，分別為：Accuracy Score, Precision, Recall, F1 Score，會選擇這四個測量方式是為了與文獻中的模型比較，文獻中模型的呈現結果也是使用這四個數值測量模型結果。
- 採用平台：Google Colab (Python)

4、 分析結果 (Analysis results)

- 實驗結果（需有圖表，以及針對圖表的闡釋）

AUROC 為 Presentation 要求附上，因此報告有補，但因原始參考資料沒有就標示為 no data。

- Apriori minimum support 選擇

  minimum support 大小會影響所拿來訓練的 frequency itemsets 數量(每一個 frequency itemsets 為一個 input column)

  下表以 KNN 為例，做了以下測試選擇了可以得到最高結果的 minimum support。由下表可知，minimum support=0.8 時，所挖掘出來的頻繁集過少，影響後續訓練指標(F1 Score、Accuracy、Precision、Recall、AUROC)，而 minimum support=0.5 時，雖然挖出很多頻繁集但有 overfitting 問題導致具體結果反而比 minimum support=0.8 時低，但是 auroc 卻剛好相反，此問題會在針對**實驗結果提出看法，並進行細項討論**說明。

Table 1. Different metrics of minimum support

minimum support	frequency itemset 數量	F1 Score	Accuracy	Precision	Recall	AUROC
0.5	57	74.11%	72.00%	81.05%	72.00%	49.29%
0.7	47	85.36%	85.33%	85.46%	85.33%	45.18%
0.75	47	85.36%	85.33%	85.46%	85.33%	45.18%
0.8	9	80.93%	81.33%	81.42%	81.33%	46.88%

- 所有 Model 分析實驗結果

  baseline 選用論文中的 report 得來的結果，為 KNN 加 The Genetic Algorithm. 由下表可以得到

  - F1 Score、Accuracy、Precision、Recall：Random Forest 和 XGBoost 和 Decision Tree 的上列指標完全一致，均達到 90.67% 左右，表現最佳。
  - Random Forest 的 AUROC 值最高，達到 96.41%，略高於 XGBoost 的 96.04%、Decision Tree 的 89.11%。

- SVM 在 F1 Score、Accuracy、Precision、Recall 上表現低於上述最好的 model，但 AUROC 達 95.98%比 Decision Tree 高，顯示其模型的分類效果依然接近最優。
- KNN 的 F1 Score、Accuracy、Precision、Recall 均約為 85%，AUROC 僅 45.18%，說明模型在區分不同類別時效果較差。

Table 2. Different metrics of different model

	F1 Score	Accuracy	Precision	Recall	AUROC
Decision Tree	90.58%	90.67%	90.55%	90.67%	89.11%
Random Forest	90.58%	90.67%	90.55%	90.67%	96.41%
XGBoost	90.58%	90.67%	90.55%	90.67%	96.04%
SVM	89.18%	89.33%	90.00%	89.33%	95.98%
KNN	85.36%	85.33%	85.46%	85.33%	45.18%
Baseline (KNN with GA)	83.21%	83.19%	83.42%	83.18%	No data

- 針對實驗結果提出看法，並進行細項討論
  - SVM 的 F1 Score、Accuracy、Precision、Recall 低於 Decision tree 但是 AUROC 卻相反:

    SVM 使用超平面來找到不同類別之間的最大間隔，專注於邊界附近的少數關鍵樣本，因為它能在不同類別之間找到較平滑且準確的決策邊界，特別適合處理少數類別的數據。

- KNN AUROC 指標較其他模型偏低:

  可能是因為 KNN 依靠距離去分類，但在 apriori 過程中已經以有無符合頻繁集(0 或 1)取代連續資料，雖然有考慮大概的 range，但因此演算法對於距離最為敏感，所以才會導致 AUROC 明顯偏低。

- KNN AUROC 與其他指標對於 minimum support 表現相反原因:

  較低的 minimum support，有較多維度資料(47 column)，對於 KNN 來說維度高表現反而降低。而在 minimum support = 0.5 (57 column) AUROC 較高的原因可能是因為較多的數據可以反應較多的距離資訊，因 KNN 對於距離資訊最為敏感。

5、 過程中遭遇的挑戰以及總結(Discussion and Conclusion)

- 描述實作 Project 中遇到的難題，以及對應之解決方法

1. 資料集樣本數量太少：

   Sol. 沒有特別處理，這個問題只能再針對類似題材蒐集更多資料才能解決。

2. 樣本 class 比例不平衡：

   Sol. 使用 SMOTE 解決資料不平衡的問題。

3. 資料集 BMI Column 的 label 沒有統一，”Normal”和”Normal Weight”都是同一個意思：

   在 Data preprocessing 時統一處理成”Normal”

- 針對此 Project 進行總結

本 project 透過 Sleep Health and Lifestyle Dataset， 探討影響睡眠障礙的關鍵因素，並建立模型進行預測。

資料包含 374 筆樣本與 13 個欄位，涵蓋睡眠時長、壓力指數、BMI 分類等多元資訊。處理流程包括 Data Pre-process、feature importance 及使用 Apriori algorithms 找到 association rules，以分析失眠與睡眠呼吸中止症的相關因素。模型訓練採用 Decision Tree、Random Forest、XGBoost 等方法，Evaluation metrics 包括 Accuracy、Precision、Recall 及 F1 Score。實驗結果顯示，隨機森林與 XGBoost 表現最佳，平均效能達 90.67%。重要的影響因素包括血壓、BMI、睡眠時長與年齡。此外，Apriori 分析揭示 BMI 正常者較無睡眠障礙疑慮，BMI 過重者易患失眠或睡眠呼吸中止症。本次 project 成功驗證模型在小資料集下的準確性，為改善睡眠健康提供有價值的參考建議。

6、 參考文獻 (Reference)

Applying Machine Learning Algorithms for the Classification of Sleep Disorders

https://www.thensf.org/sleep-facts-and-statistics/

https://www.researchgate.net/publication/372975033_Classification_of_Sleep_Disorders_Using_Random_Forest_on_Sleep_Health_and_Lifestyle_Dataset

https://ieeexplore.ieee.org/document/10462120/references#references

7、 組員分工與各自執行細項 (Work distribution chart)

Table 3. 組員分工與各自執行細項

Member	Work Distribution
林慧旻	Data preprocessing, Data Analysis, Report
洪明祺	Apriori, Model - KNN, Report
李以恩	Model - Random Forest, XGBoost, SVM, Presentation
林念慈	Feature Importance, Model - Decision Tree, Report