

金融大數據期末報告

睡眠品質分析

第三組

資財碩一 112AB8002 洪于喬

資財碩一 112AB8019 林寧

資財碩一 112AB8034 李明禧

資財碩一 112AB8408 范氏明奎

電子四甲 109360148 曾偉為

1. 研究目的

本研究的主要目的是找出提升睡眠品質的方法，並識別與睡眠品質直接相關的關鍵因素。我們關注的主要問題包括：哪些因素會直接影響睡眠品質？如何通過調整這些因素來提高睡眠品質？

為此我們將研究層面分為四層，首先是基本睡眠品質分析，探討性別、年齡、職業、壓力程度對睡眠品質的影響。其次是生活方式對睡眠的影響，分析身體活動量、步數、BMI 對睡眠品質的影響。第三是心血管健康對睡眠的影響，分析血壓、心率與 BMI 與睡眠品質的關係和影響。最後是整體睡眠狀況分析，全面分析顯著因素對睡眠品質的影響。

2. 資料來源

本研究使用的資料集來自 Kaggle 網站的公開資料集——Sleep Health and Lifestyle Dataset。該資料集涵蓋了與睡眠和日常習慣相關的廣泛變數，共有 13 個欄位和 374 條樣本。主要變數包括：

欄位名稱	說明
Person ID (人員 ID)	An identifier for each individual.
Gender (性別)	The gender of the person (Male/Female).
Age (年齡)	The age of the person in years.
Occupation (職業)	The occupation or profession of the person.
Sleep Duration (睡眠時間)	The number of hours the person sleeps per day.
Quality of Sleep (睡眠品質)	A subjective rating of the quality of sleep, ranging from 1 to 10.
Physical Activity Level (身體活動程度)	The number of minutes the person engages in physical activity daily.
Stress Level (壓力程度)	A subjective rating of the stress level experienced by the person, ranging from 1 to 10.
BMI Category (BMI 類別)	The BMI category of the person (e.g., Underweight, Normal, Overweight).
Blood Pressure (血壓)	The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.
Heart Rate (心率)	The resting heart rate of the person in beats per minute.
Daily Steps (每日步數)	The number of steps the person takes per day.

Sleep Disorder (睡眠障礙)	The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).
-----------------------	--

3. 資料預處理

首先進行缺失值填補、和重複紀錄的刪除，實驗檢查數據集無缺漏值，也無重複紀錄，接者進行數據格式化，確保數據遵循一致的格式和標準，第三進行數據離散化，將連續數據轉換為分類數據，提高模型的解釋性，第四進行類別重編碼，將部分類別合併成一類，避免稀疏數據問題，最後進行數據標準化，將不同特徵轉換到相同的量綱，使得數據具有可比性。

1. 填補缺失值、刪除重複紀錄：檢查無缺漏值、無重複紀錄
2. 數據格式化：BMI Category 的 Normal 和 Normal Weight 具有相似的含義，將 Normal Weight 轉換為 Normal；Sleep Disorder 中有三個類別，其中 None 表示沒有睡眠障礙，將 None 轉換成 Normal。
3. 數據離散化：Blood Pressure 的格式為（收縮壓/舒張壓），提取出收縮壓 Systolic 和舒張壓 Diastolic，並透過判斷式，分成 Optimal、Normal、Hypertension 三個類別，放在新的欄位 Blood Pressure Category；Daily Steps 分布不均，為了緩解極端值的影響，進行分箱處理，分成四個區間；Physical Activity Level 分布不均，為了緩解極端值的影響，進行分箱處理，分成三個區間。

4. 類別重編碼：Occupation 的每一種職業採樣不均，對於頻率較低的類別，將近似的合併，最後留下四個類別：醫護、科研、業務、其他；Sleep Quality 的數值不全，睡眠品質採樣只有 4, 5, 6, 7, 8, 9，為了不影響模型訓練，將 Sleep Quality 重編碼為三個類別：1, 2, 3。

5. 數據標準化：將所有數值型欄位進行最小-最大標準化，將數據縮放到[0, 1]。

4. 資料敘述統計

1. 單變量敘述統計

#	Column	Non-Null Count	Dtype
0	Person ID	374 non-null	int64
1	Gender	374 non-null	object
2	Age	374 non-null	int64
3	Occupation	374 non-null	object
4	Sleep Duration	374 non-null	float64
5	Quality of Sleep	374 non-null	int64
6	Physical Activity Level	374 non-null	int64
7	Stress Level	374 non-null	int64
8	BMI Category	374 non-null	object
9	Blood Pressure	374 non-null	object
10	Heart Rate	374 non-null	int64
11	Daily Steps	374 non-null	int64
12	Sleep Disorder	374 non-null	object

圖 1 數據預處理前

#	Column	Non-Null Count	Dtype
0	Person ID	374 non-null	int64
1	Gender	374 non-null	object
2	Age	374 non-null	int64
3	Occupation	374 non-null	object
4	Sleep_Duration	374 non-null	float64
5	Sleep_Quality	374 non-null	object
6	Stress_Level	374 non-null	int64
7	BMI_Category	374 non-null	object
8	Heart_Rate	374 non-null	int64
9	Sleep_Disorder	374 non-null	object
10	BP_Category	374 non-null	object
11	Daily_Steps_Group	374 non-null	object
12	Physical_Activity_Group	374 non-null	object

圖 2 數據預處理後

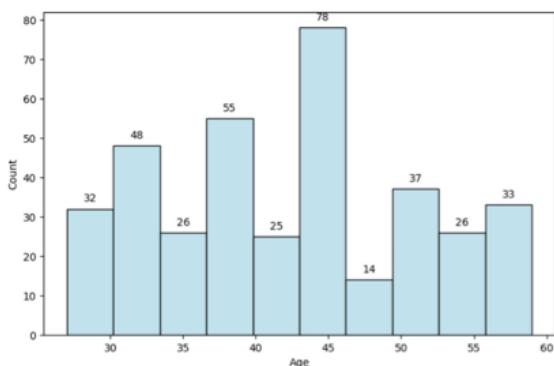


圖 3 Age

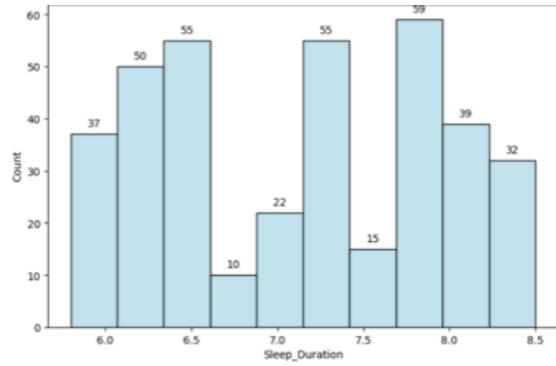


圖 4 Sleep Duration

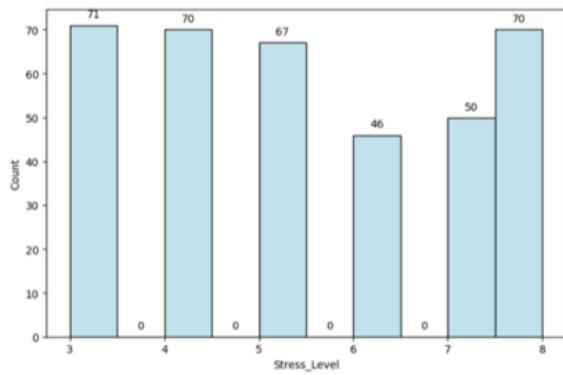


圖 5 Stress Level

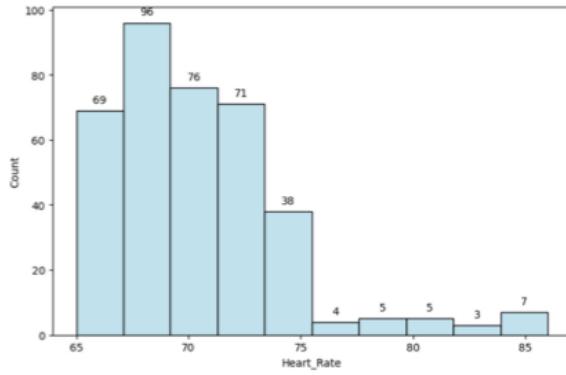


圖 6 Heart Rate

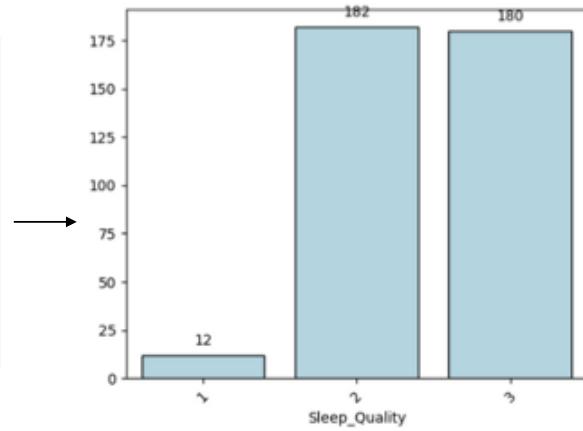
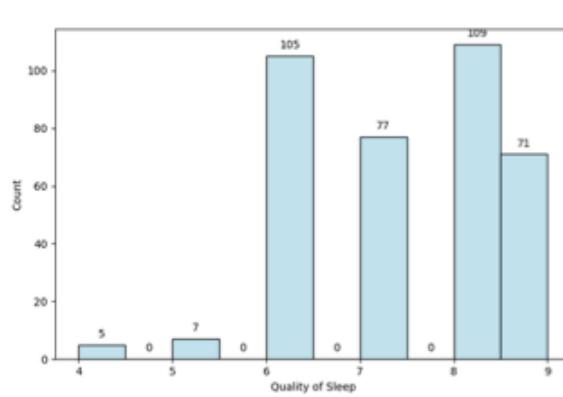


圖 7 Sleep Quality

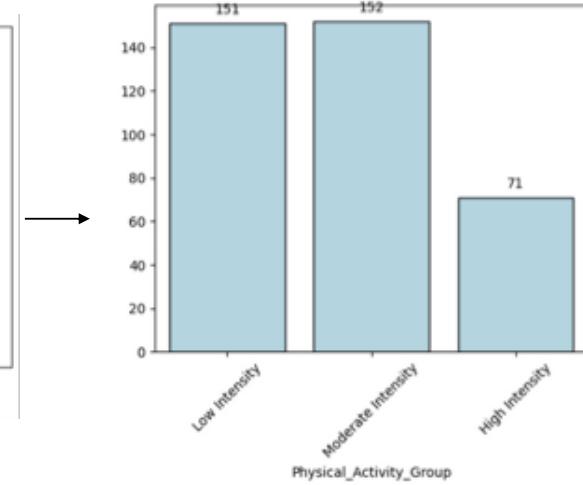
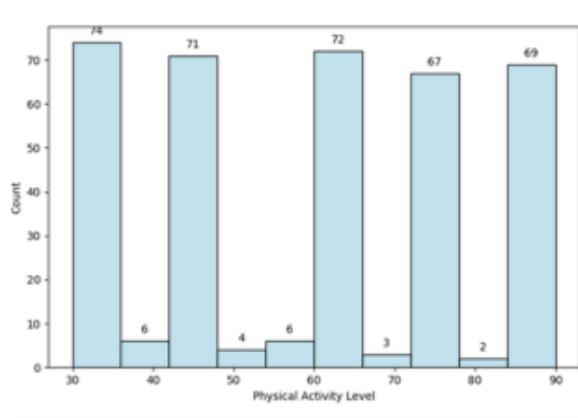


圖 8 Physical Activity

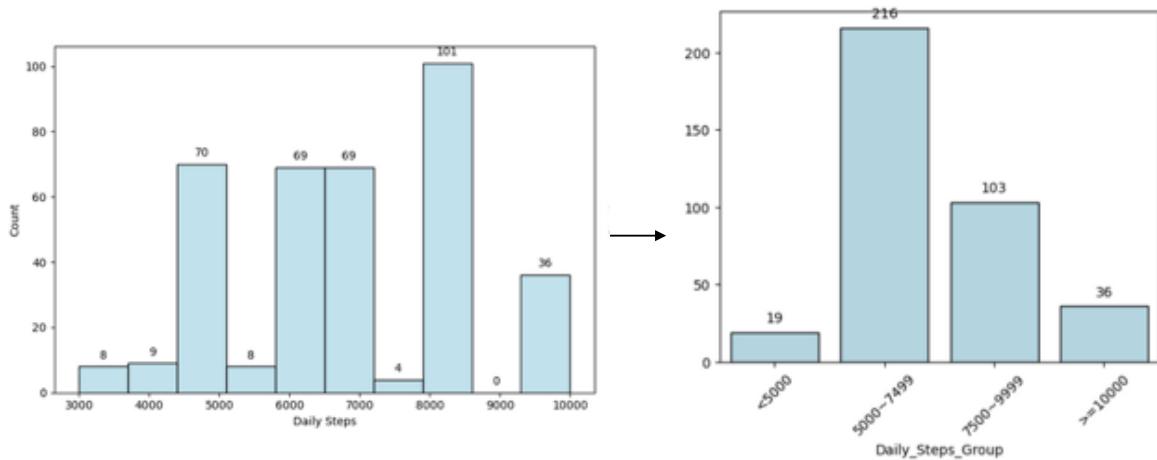


圖 9 Daily Steps

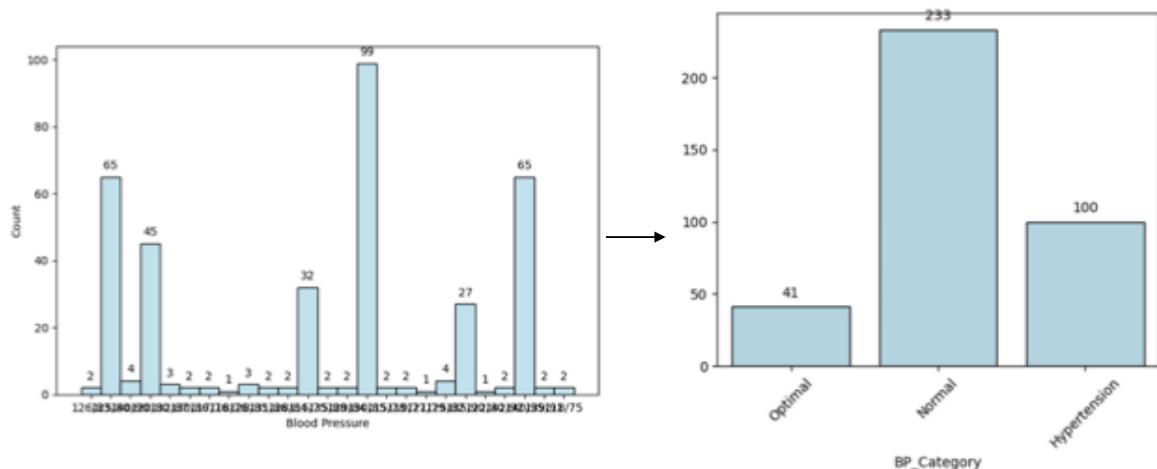


圖 10 Blood Pressure

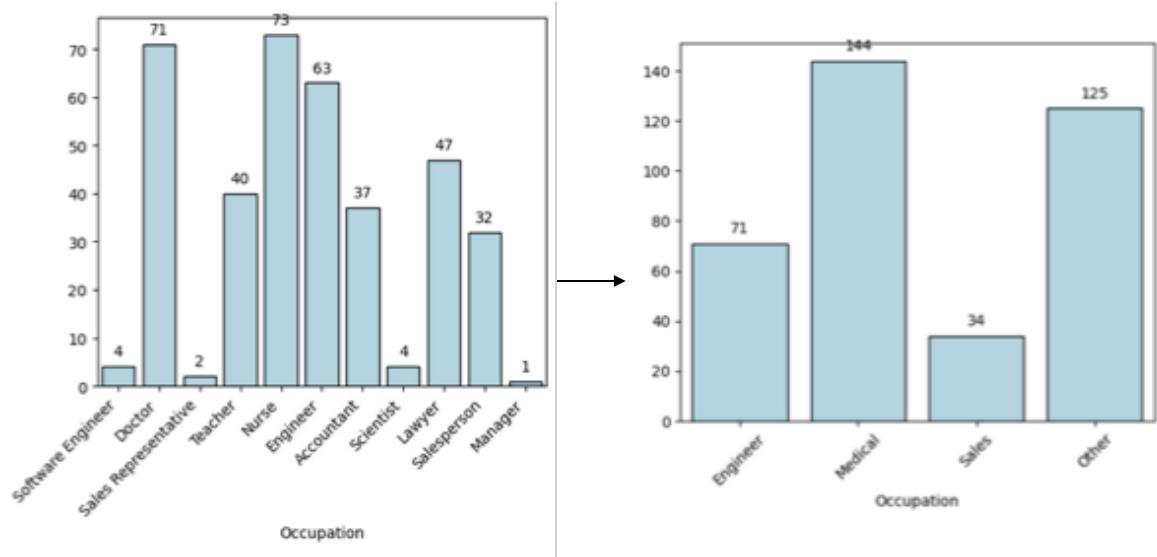


圖 11 Occupation

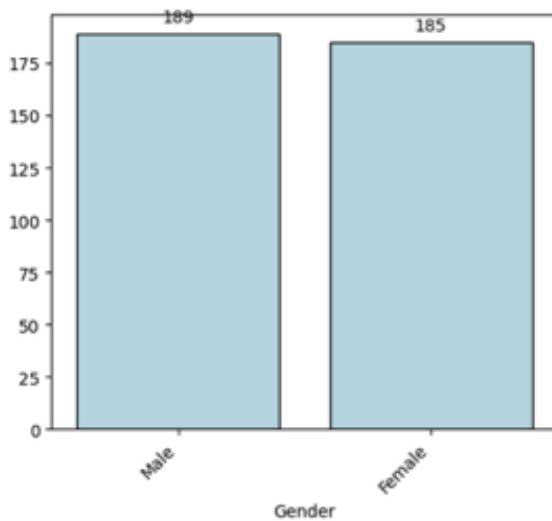


圖 12 Gender

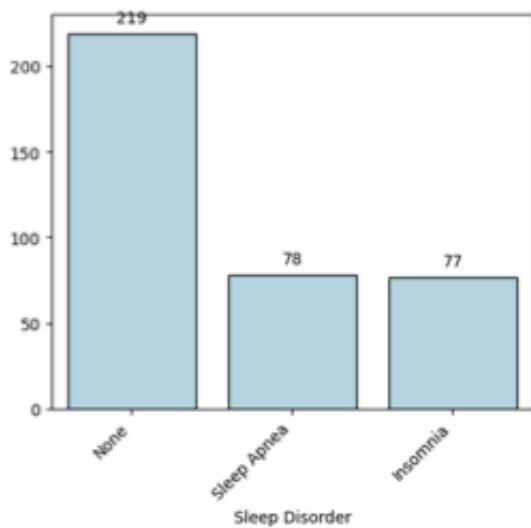


圖 13 Sleep Disorder

2. 相關性分析

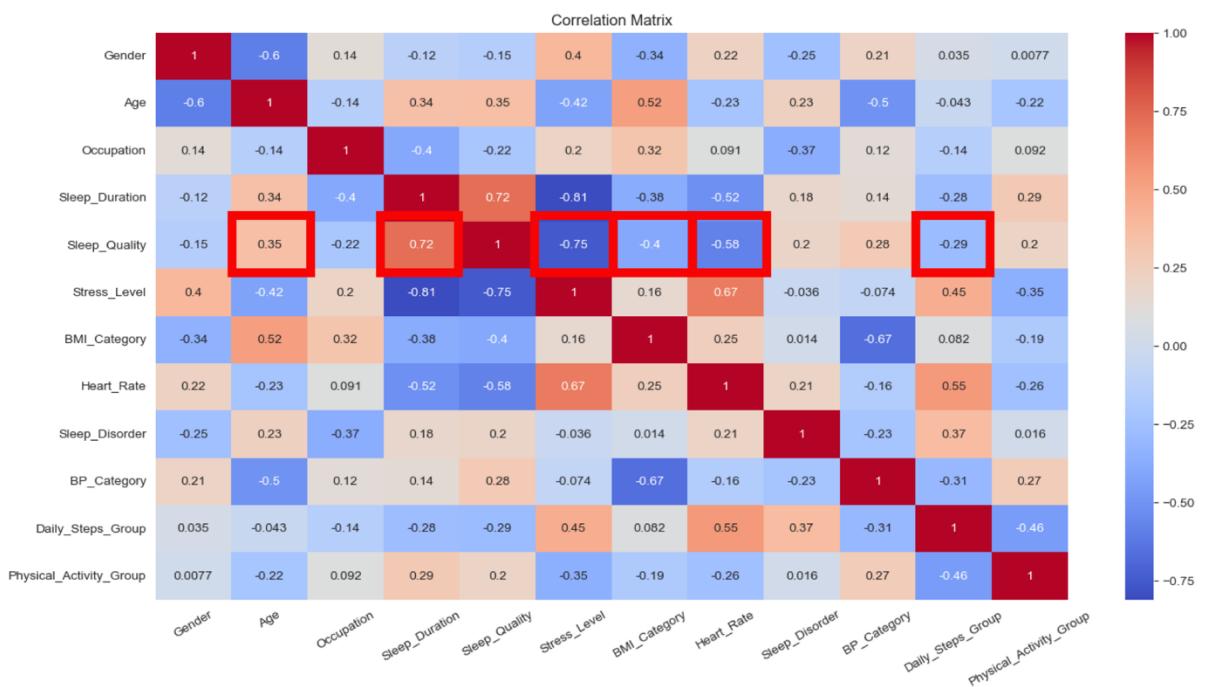


圖 14 相關係數矩陣

從相關係數矩陣中，發現與 Sleep Quality 相關係數絕對值較大的前六個變量，從大到小分別是，Stress Level、Sleep Duration、Heart Rate、BMI Category、Age、

Daily Steps Group，其中 Sleep Duration 與 Sleep Quality 呈強烈正相關，因此在做特徵選擇時會將 Sleep Duration 變量排除。

針對 Stress Level、Heart Rate、BMI Category、Age、Daily Steps Group 這五個變量做多變量敘述統計：

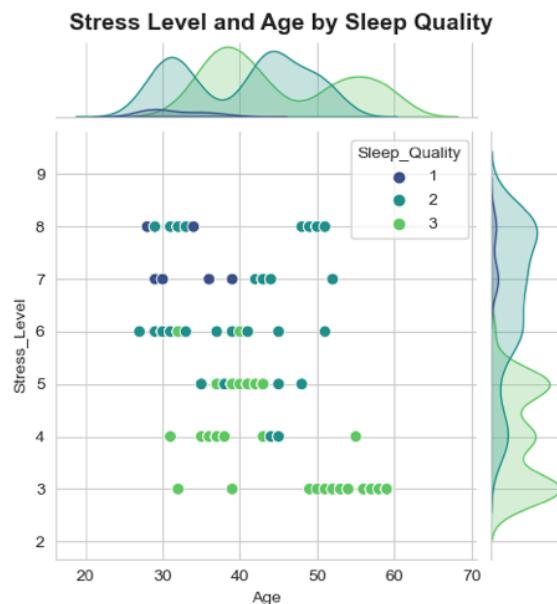


圖 15 壓力程度與年齡和睡眠品質的分佈



圖 16 壓力程度與心率和睡眠品質的分佈

壓力程度越大，睡眠品質越差；心率越低，睡眠品質越高；壓力、心率與睡眠品質呈顯著負相關。

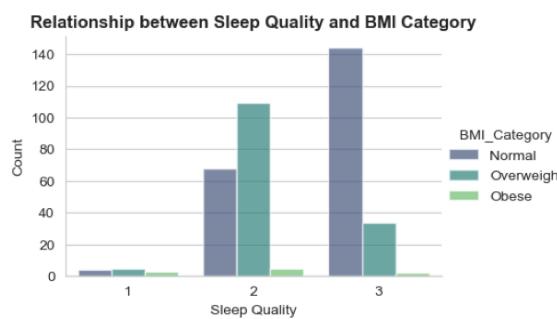


圖 17 睡眠品質與 BMI 的關係

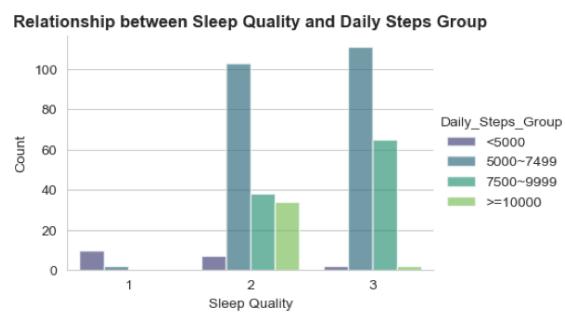


圖 18 睡眠品質與 BMI 的關係

BMI 落在正常範圍的群體，睡眠品質高的人較多；而 BMI 落在過重以上的群體，睡眠品質普通的人較多。每日步數達 7500 的群體，睡眠品質都偏高；而每日步數不到 5000 的群體，睡眠品質低的人較多。

5. 資料分析方法

通過機器學習算法的特徵重要性分析，來識別出對睡眠質量有顯著影響的生理、心理和生活方式因素。為此我們選擇了三種分類算法分別是 Random Forest Classifier 隨機森林分類、Decision Tree Classifier 決策樹分類、K Nearest Neighbors Classifier K 最近鄰分類。

Random Forest Classifier 隨機森林分類：

隨機森林分類器是一種集成學習方法，通過結合多個決策樹來提高分類的準確性和穩定性，隨機森林使用多棵決策樹進行分類，每棵樹都是在不同的隨機子樣本和特徵子集上訓練的。通過結合多棵決策樹的結果，隨機森林能夠降低單一決策樹的過擬合問題，提高模型的泛化能力。

首先構建多棵樹，從訓練數據集中隨機選擇有放回的在構建每棵決策樹時，接著隨機特徵選擇，在構建每棵決策樹時，對於每個節點只從所有特徵中隨機選擇一部分特徵進行分割，後集成結果，對於每個節點只從所有特徵中隨機選擇一部分特徵進行分割。子樣本（bootstrap samples）來訓練每棵決策樹。

Decision Tree Classifier 決策樹分類：

決策樹分類器是一種基於樹狀結構的分類模型，每個內部節點表示一個特徵（或屬性）的測試，每個分支表示測試結果，而每個葉子節點則表示類標籤（或分佈）。易於理解和解釋，模型結構直觀，決策過程容易解釋。是一種非參數模型，不需要對數據的分佈做任何假設，能夠處理多種數據類型，且能夠處理數值型和類別型數據。

首先選擇最佳分割特徵，使用某種標準（如信息增益、基尼係數等）選擇最佳特徵來分割數據，接著遞歸構建樹，將數據集分割成子集，對每個子集重複上述過程，直到滿足停止條件（如所有樣本屬於同一類別或不再有可分割的特徵），最後樹的剪枝，通過減少分支來簡化樹結構，防止過擬合。

K Nearest Neighbors Classifier K 最近鄰分類

K 最近鄰分類器（KNN）是一種基於實例的學習方法，不使用顯式的訓練過程，而是根據距離度量來進行分類。屬於懶惰學習，沒有顯式的訓練階段，分類時才使用訓練數據，算法是基於距離度量，通常使用歐氏距離（或其他距離度量）來計算樣本之間的距離。

首先計算距離，對於待分類樣本，計算它與訓練數據集中所有樣本之間的距離，接著選擇鄰居，找出距離最近的 K 個鄰居，最後投票決策：根據這 K 個鄰居的類標籤進行投票，決定待分類樣本的類別（多數決）。

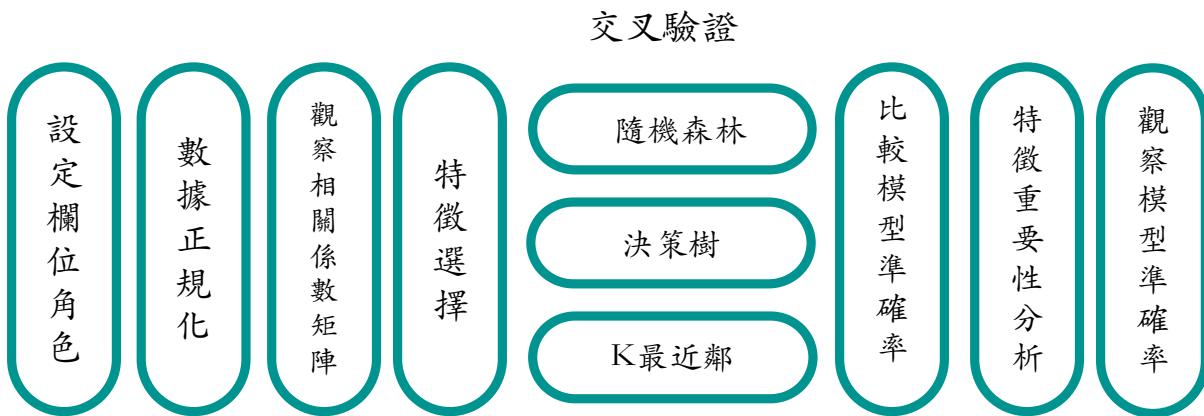


圖 19 分析方法流程圖

1. 預測睡眠品質

表 1 Random Forest 隨機森林訓練表現

	Random Forest				
整體 (全部變量)	98.94%				
	<small>accuracy: 98.94% +/- 1.37% (micro average: 98.93%)</small>				
		true 1	true 2	true 3	class precision
	pred. 1	12	1	0	92.31%
	pred. 2	0	180	2	98.90%
	pred. 3	0	1	178	99.44%
	class recall	100.00%	98.90%	98.89%	
基本 (壓力水平、職業、年齡、性別)	94.92%				
	<small>accuracy: 94.92% +/- 2.67% (micro average: 94.92%)</small>				
		true 1	true 2	true 3	class precision
	pred. 1	3	0	0	100.00%
	pred. 2	9	176	4	93.12%
	pred. 3	0	6	176	96.70%
	class recall	25.00%	96.70%	97.78%	
生活方式 (身體活動量、每日步數、BMI)	80.50%				
	<small>accuracy: 80.50% +/- 6.50% (micro average: 80.48%)</small>				
		true 1	true 2	true 3	class precision
	pred. 1	7	3	0	70.00%
	pred. 2	5	146	32	79.78%
	pred. 3	0	33	148	81.77%
	class recall	58.33%	80.22%	82.22%	
心血管	88.24%				

(血壓、心律、BMI)		accuracy: 88.24% +/- 2.86% (micro average: 88.24%)				
		true 1		true 2	true 3	class precision
pred. 1		12		0	0	100.00%
pred. 2		0		170	32	84.16%
pred. 3		0		12	148	92.50%
class recall		100.00%		93.41%	82.22%	

表 2 Decision Tree 決策樹訓練表現

	Decision Tree				
	96.53%				
整體 (全部變量)	accuracy: 96.53% +/- 3.09% (micro average: 96.52%)				
基本 (壓力水平、職業、年齡、性別)	94.64%				
	accuracy: 94.64% +/- 5.09% (micro average: 94.65%)				
生活方式 (身體活動量、每日步數、BMI)	79.68%				
	accuracy: 79.68% +/- 6.26% (micro average: 79.68%)				
心血管 (血壓、心律、BMI)	85.53%				
	accuracy: 85.53% +/- 5.63% (micro average: 85.56%)				

表 3 Random Forest 隨機森林訓練表現

	KNN			
整體	94.15%			

(全部變量)	accuracy: 94.15% +/- 3.88% (micro average: 94.12%)				
		true 1	true 2	true 3	class precision
	pred. 1	9	7	0	56.25%
	pred. 2	3	171	8	93.96%
	pred. 3	0	4	172	97.73%
基本 (壓力水平、職業、年齡、性別)	class recall	75.00%	0	93.96%	95.56%
	93.87%				
	accuracy: 93.87% +/- 3.29% (micro average: 93.85%)				
		true 1	true 2	true 3	class precision
	pred. 1	4	1	0	80.00%
生活方式 (身體活動量、每日步數、BMI)	pred. 2	7	172	5	93.48%
	pred. 3	1	9	175	94.59%
	class recall	33.33%	94.51%	97.22%	
	70.89%				
	accuracy: 70.89% +/- 5.30% (micro average: 70.86%)				
心血管 (血壓、心律、BMI)		true 1	true 2	true 3	class precision
	pred. 1	10	5	2	58.82%
	pred. 2	2	175	98	63.64%
	pred. 3	0	2	80	97.56%
	class recall	83.33%	96.15%	44.44%	
	86.38%				
	accuracy: 86.38% +/- 6.03% (micro average: 86.36%)				
		true 1	true 2	true 3	class precision
	pred. 1	7	3	2	58.33%
	pred. 2	5	178	40	79.82%
	pred. 3	0	1	138	99.28%
	class recall	58.33%	97.80%	76.67%	

從三種模型的訓練表現結果中，發現 Random Forest 的模型表現最佳，因此本研究針對 Random Forest 做特徵重要性分析，發現前 5 大重要的特徵分別是：Sleep Duration、Stress Level、Age、BP Category，其中 Sleep Duration 與 Sleep Quality 有強烈的正相關，所以將此變量排除，做特徵選擇後，留下了四個特徵：Stress Level、Heart Rate、Age、BP Category，再次使用隨機森林訓練，結果準確率高達 97.87！

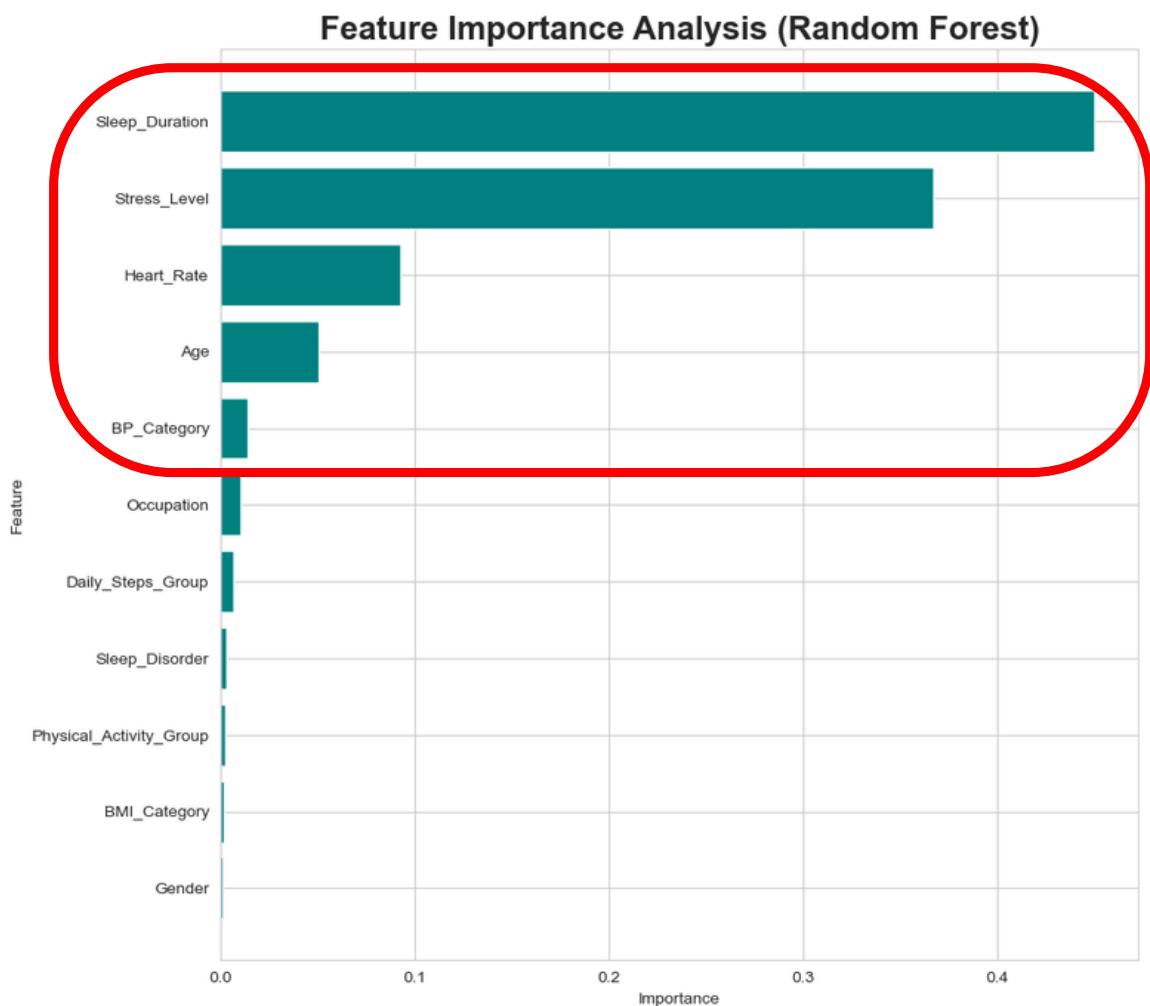


圖 20 睡眠品質的特徵重要性分析

accuracy: 97.87% +/- 2.08% (micro average: 97.86%)

	true 1	true 2	true 3	class precision
pred. 1	12	1	0	92.31%
pred. 2	0	177	3	98.33%
pred. 3	0	4	177	97.79%
class recall	100.00%	97.25%	98.33%	

圖 21 特徵選擇過後的混和矩陣

結果顯示 Sleep_Quality = 1 召回率達 100.00%，精確度只有 92.31%，可能原因是樣本數較少。Sleep_Quality = 2 的召回率達 97.25%，精確度達 98.33%，表現不錯。Sleep_Quality = 3 的召回率達 98.33%，精確度達 97.79%，表現相當良好。

2. 睡眠品質提升分析

本研究的角色設定為“台灣睡眠醫學學會研究人員”，因此致力於幫助民眾提升睡眠品質！從模型特徵重要性分析中，識別出影響睡眠品質的三大關鍵因素，因此本研究希望透過調整關鍵因素，來提升睡眠品質。

- | | |
|-----------------------|-----------------|
| 1 睡眠時長 Sleep Duration | 較容易直接調整 |
| 2 壓力程度 Stress Level | 較不易直接調整，但可以間接調整 |
| 3 心率 Heart Rate | 較不易直接調整，但可以間接調整 |

圖 22 關鍵因素

為了達到提升睡眠品質的目的，我們思考在個人因素（性別、年齡、職業）不變的情況下，若想要提升睡眠品質到達某一程度，該如何透過改變睡眠時長、壓力程度、心率來提升睡眠品質？

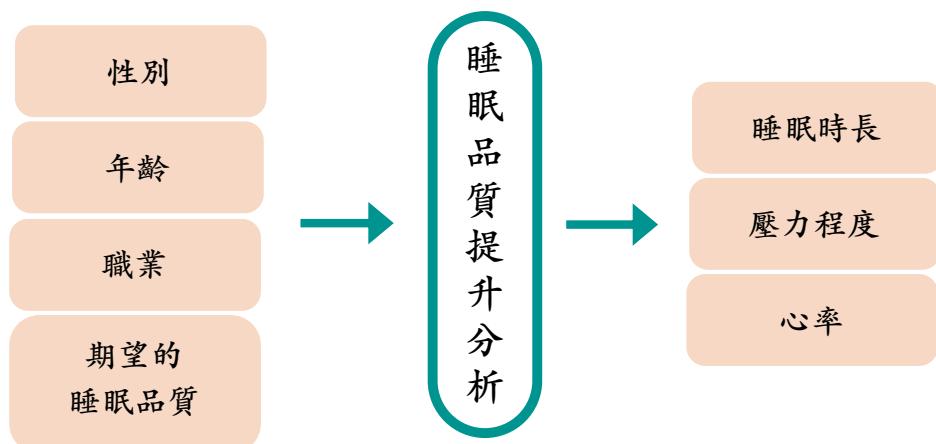


圖 23 睡眠品質提升分析

所以針對關鍵因素，我們進一步分析預測找出能夠影響關鍵因素的次要因素，發現每個因素之間環環相扣的關係，才能更好的達到睡眠品質提升的效果！

3. 睡眠時長分析

經過隨機森林的訓練後，本研究得出睡眠時長模型的準確率為 83.95%，特徵重要性分析發現，影響睡眠時長的關鍵因素分別為：Sleep Disorder、Age、Stress Level、BMI Category、Occupation。

accuracy: 83.95% +/- 4.91% (micro average: 83.96%)							
	true ~6	true 6-6.5	true 6.5-7	true 7.5-8	true 7-7.5	true 8~	class precision
pred. ~6	30	18	0	6	0	0	55.56%
pred. 6-6.5	3	51	0	0	1	0	92.73%
pred. 6.5-7	0	12	32	0	2	0	69.57%
pred. 7.5-8	4	4	0	69	2	0	87.34%
pred. 7-7.5	0	0	1	2	74	0	96.10%
pred. 8~	0	0	0	5	0	58	92.06%
class recall	81.08%	60.00%	96.97%	84.15%	93.67%	100.00%	

圖 24 睡眠時長模型表現

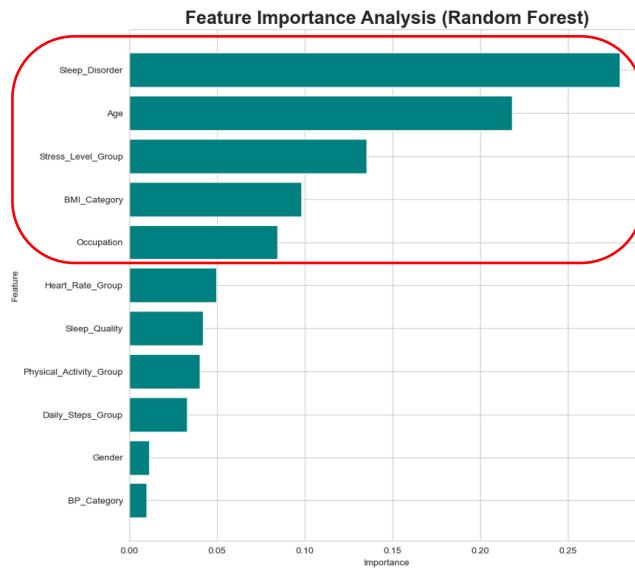


圖 25 睡眠時長的特徵重要性分析

4. 壓力程度分析

經過隨機森林的訓練後，本研究得出壓力程度模型的準確率為 91.96%，特徵重要性分析發現，影響睡眠時長的關鍵因素分別為：Heart Rate、Sleep Quality、Gender、Age、Physical Activity，從中發現壓力程度與 Sleep Quality 是互相影響的雙向關係！

accuracy: 91.96% +/- 3.15% (micro average: 91.98%)

	true 4	true 6	true 5	true 2	true 1	true 3	class precision
pred. 4	31	6	0	0	0	1	81.58%
pred. 6	9	64	1	0	0	0	86.49%
pred. 5	1	0	48	0	0	0	97.96%
pred. 2	2	0	0	69	2	3	90.79%
pred. 1	1	0	0	0	69	0	98.57%
pred. 3	2	0	1	1	0	63	94.03%
class recall	67.39%	91.43%	96.00%	98.57%	97.18%	94.03%	

圖 26 壓力程度模型表現

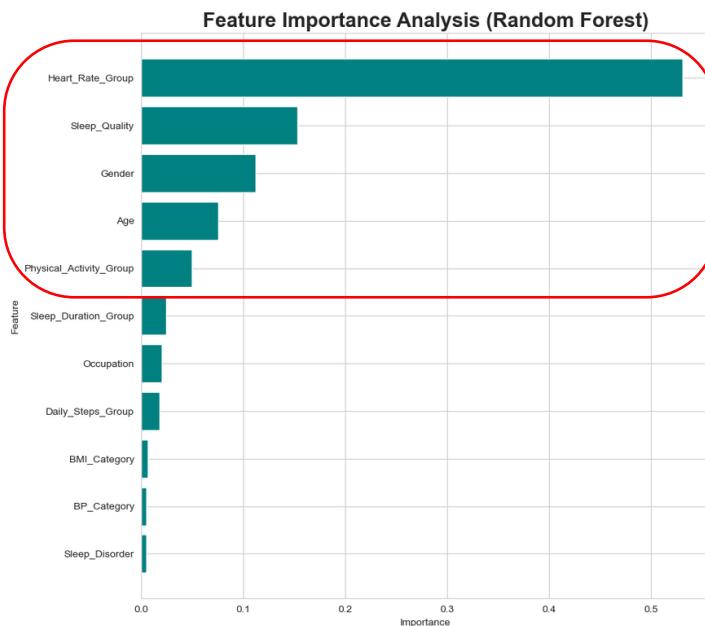


圖 27 壓力程度的特徵重要性分析

5. 心率分析

經過隨機森林的訓練後，本研究得出心律模型的準確率為 94.38%，特徵重要性分析發現，影響睡眠時長的關鍵因素分別為：Stress Level、Age、Daily Steps、Sleep Duration、BP Category，從中發現壓力程度與 Stress Level 是互相影響的雙向關係！

accuracy: 94.38% +/- 4.85% (micro average: 94.39%)

	true 75~80	true 70~75	true 80~85	true ~70	true 85~	class precision
pred. 75~80	8	0	0	1	0	88.89%
pred. 70~75	2	101	1	3	0	94.39%
pred. 80~85	1	1	7	2	0	63.64%
pred. ~70	1	7	2	235	0	95.92%
pred. 85~	0	0	0	0	2	100.00%
class recall	66.67%	92.66%	70.00%	97.51%	100.00%	

圖 28 心率模型表現

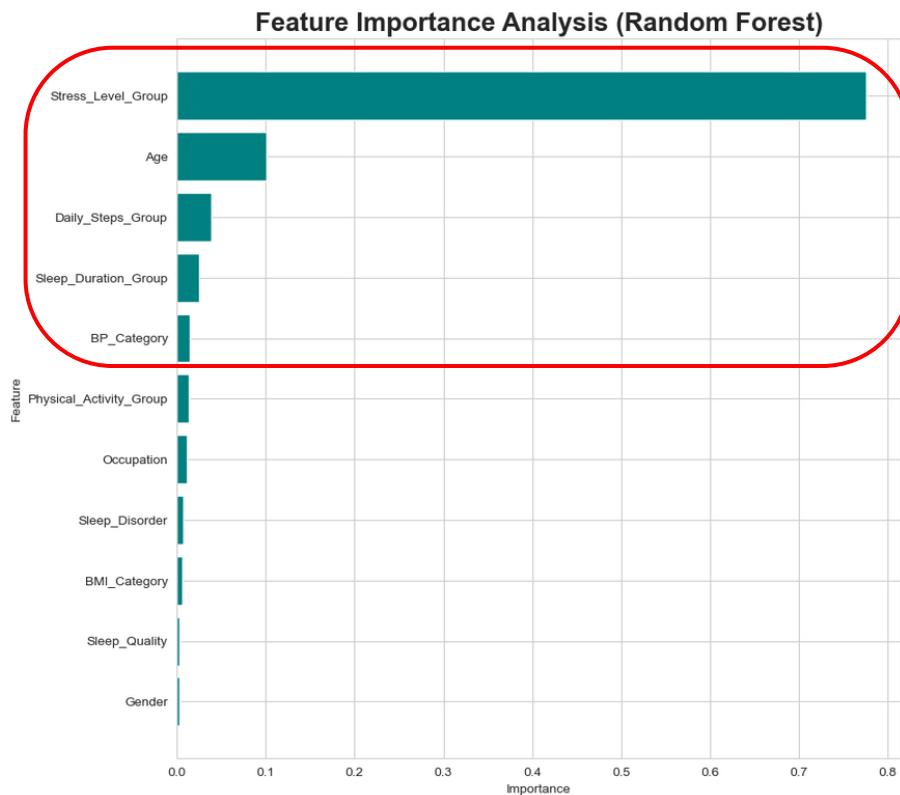


圖 29 心率的特徵重要性分析

6. 文獻支持

Bilal A. Chaudhry , et al. , The Relationship between Sleep Duration and Metabolic Syndrome Severity Scores in Emerging Adults. (2023) 提到研究表明短睡眠時間 (<7 小時) 和長睡眠時間 (>9 小時) 都與較高的代謝症候群嚴重程度評分相關，這表明最佳睡眠時間對於代謝健康和整體健康至關重要。Mirjam Ekstedt , et al. , Microarousals during sleep are associated with increased levels of lipids, cortisol, and blood pressure. (2004) 顯示高壓力水平一直與較差的睡眠品質有關。壓力會激活下丘腦-垂體-腎上腺 (HPA) 軸，增加皮質醇水平，從而擾亂睡眠模式並降低睡眠效率。Yongbin Li , et al. , Research on the relationship between physical activity, sleep quality, psychological resilience, and social adaptation among Chinese college students: A cross-sectional study. (2023) 研究發現體育活動可以顯著改善大學生的心理恢復力和社會適應，從而改善睡眠質量，表明參與體育活動可能有助於減少這個人群的睡眠問題。Hanne K J Gonnissen , et al. , Sleep duration, sleep quality and body weight: Parallel developments. (2013) 研究討論了睡眠質量和體重之間的關係，強調睡眠短或受干擾與肥胖的增加有關，且青春期和成年期的 BMI 指數變化與睡眠時長的變化相反相關，表明更好的睡眠可能有助於管理體重。Heart rate variability, sleep and sleep disorders. (2012) 研究調查了失眠患者的心率變異性 (HRV) 與睡眠品質之間的關係。結論是較低的 HRV (表示較高的心率) 與較差的睡眠品質相關。Amirreza Sajjadieh , et al. , The Association of Sleep Duration and Quality with Heart Rate Variability and Blood Pressure. (2020) 這項研究評估了青少年心率變異性和睡眠效率之間的關聯。研究結果表明，較高的心率與較低的睡眠效率有關，導致睡眠品質較差。

6. 分析結果與討論

分析結果顯示，影響睡眠品質的三大關鍵因素為：睡眠時長、壓力程度、心率，且三者之間存在互相影響的關係。另外分析結果也顯示睡眠品質會影響壓力程度。

本研究提出兩個方法改善睡眠品質，首先是直接調整睡眠的時間長度，不要過短或過長，在一個適合的範圍內；其次是透過運動（身體活動）調整身體素質（BMI）和心血管健康，進而排解壓力、穩定心率，才能夠進一步提高睡眠品質。

本研究發現的潛在的限制和假設，首先環境因素差異，因此身體基本素質會有所不同，其次職業不平均，可能無法涵蓋各種職業，第三樣本數偏少，資料樣本數不夠多，無法代表整體，最後是樣本數偏少，資料樣本數不夠多，無法代表整體！

7. 結論

影響睡眠品質的因素有很多，每個因素之間也都環環相扣，本研究發現，為了提高人們的睡眠品質，在固定因素(年齡、性別、職業)無法改變的情況下，透過調整自身生活模式，進而改善心血管健康狀況，能夠提升個人的睡眠品質。

本研究對未來研究的建議是，首先要對諮詢者進行大量問卷調查，接著根據收集到的數值進行分析，並且提出個性化的改善建議和方法，最後評估調整後的改善情況並持續追蹤受測者的睡眠情況！

8. 參考資料

1. Kaggle , Sleep Health and Lifestyle Dataset :
<https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>
(2024)
2. PubMed , The Relationship between Sleep Duration and Metabolic Syndrome Severity Scores in Emerging Adults :
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9965711/> (2023)
3. PubMed , Microarousals during sleep are associated with increased levels of lipids, cortisol, and blood pressure : <https://pubmed.ncbi.nlm.nih.gov/15564359/> (2004)
4. PubMed , Research on the relationship between physical activity, sleep quality, psychological resilience, and social adaptation among Chinese college students: A cross-sectional study : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9950505/>
(2023)
5. PubMed , Sleep duration, sleep quality and body weight: parallel developments :
<https://pubmed.ncbi.nlm.nih.gov/23643826/> (2013)
6. ScienceDirect , Heart rate variability, sleep and sleep disorders :
<https://www.sciencedirect.com/science/article/pii/S1087079211000293> (2012)
7. PubMed , The Association of Sleep Duration and Quality with Heart Rate Variability and Blood Pressure. : <https://pubmed.ncbi.nlm.nih.gov/33262801/>
(2020)