數據科學專題報告

身體健康狀況受何影響

11011142 曾鈺涵、11011209 謝佳璇

一、資料集介紹

資料庫名稱為全國健康老化民意調查(NPHA)。此資料集是來自「Uc Irvine Machine Learning Repository 網站」，作者是 Malani、Preeti N、Kullgren、Jeffrey、Solway 和 Erica 在 2017 年發表於大學間政治和社會研究聯盟。

創建全國健康老化民意調查資料集的目的是收集有關影響 50 歲及以上美國人的健康、醫療保健和衛生政策問題的見解。透過關注老年人及其照護者的觀點，密西根大學旨在向大眾、醫療保健提供者、政策制定者和提倡者介紹老化的各個方面。這包括健康保險、家庭組成、睡眠問題、牙科護理、處方藥和護理等主題，從而全面了解老年人口的健康相關需求和擔憂。而該資料集中每行代表一個調查受訪者。

此資料集已經進行了一些預處理，對於原始 NPHA 資料集的這個子集，我們選擇了 13 個與健康和睡眠相關的特徵來預測任務。然後刪除了所有對任何選定功能缺少回應的調查受訪者，其中沒有缺失值。以下是各欄位名稱中英對照:

1. Age:年齡
2. Physical_Health:身體健康
3. Mental_Health:精神健康
4. Dental_Health:牙科健康
5. Employment:就業
6. Stress_Keeps_Patient_from_Sleeping:壓力是否影響患者的睡眠能力
7. Medication_Keeps_Patient_from_Sleeping:藥物是否影響病人的睡眠
8. Pain_Keeps_Patient_from_Sleeping:身體疼痛是否干擾患者睡眠
9. Bathroom_Needs_Keeps_Patient_from_Sleeping:使用沐浴的需要是否影響病人的睡眠
10. Uknown_Keeps_Patient_from_Sleeping:影響患者睡眠的不明因素
11. Trouble_Sleeping:睡眠困難
12. Prescription_Sleep_Medication:處方_睡眠_藥物
13. Gender:性別

二、分析重點

使用百分比分析:

1. 比較不同身體健康狀況和精神狀況之間的關係。
2. 比較不同身體健康狀況和不明因素影響睡眠之間的關係，不明因素有可能是癌症、焦慮等等。
3. 比較不同身體健康狀況有無睡眠困難。
4. 比較不同身體健康狀況和處方睡眠藥物之間的關係。

使用 Odd Ratio 分析:

1. 身體健康狀況和是否受壓力影響睡眠的關係
2. 身體健康狀況和是否受藥物影響睡眠的關係
3. 身體健康狀況和是否受身體疼痛影響睡眠的關係
4. 身體健康狀況和是否受沐浴需求影響睡眠的關係

三、資料前處理

本資料庫中共有 715 列，其中第一列代表欄位名稱，所以訪問人數是 714 筆資料。資料前處理有分成四個部分，依序是 Missing data 的分析處理、特徵工程、特徵選擇和特徵擷取。

(A) Missing data 的分析處理

資料庫中的-1 本身是拒絕的意思，也就是 missing data。從上課所學的 Missing data Mechanism 中可得知此資料庫中的 Missing data 的種類是 Not Missing At Random (NMAR)，他是屬於沒有被記錄在資料庫內的欄位。我們使用 Complete cases analysis，將資料庫中所有含有 Missing data 欄位的訪問資料全部刪除，剩下 696 筆資料，我們接著用這些資料做後續處理。

(B) 特徵工程、選擇、擷取

特徵工程的意思是將原始資料欄位依照後續分析所需進行加工處理，在我們的資料中都是非數值欄位，我們用 label encoding，把每種分類都用數值表示出來方便做統計分析，因為 excel 的 COUNTIF 函數只能計算數值型態資料，其中的數值並沒有分大小。再用 COUNTIF 函數，得出年齡分類在第 2 組的有 696 人，而剩餘要分析的資料只有 696 筆，代表受訪者的年齡皆在 65-80 之間，在後續分析中會起不了作用，故將其刪除。而我們要做的分析當中不需要使用到就診醫師人數、種族這兩個欄位，所以也在資料前處理中先行刪除。

- EXCEL 統計各欄位的類別數量表

身體健康	精神健康	牙科健康	就業	壓力	藥物	痛苦	沐浴影響	不明原因	睡眠困難	處方睡眠藥物	性別
34	218	65	48	523	657	544	344	404	59	37	314
234	278	211	54	173	39	152	352	292	285	32	382
286	165	203	577						352	627	
121	33	126	17								
21	2	37	0								
		54									

(皆是由小到大，例如:1~6 或是 0~1)

我們使用使用 COUNTIF 函數統計出每個欄位的各類別數量，為了方便分析，我們詢問 chat GPT 如何編寫 python 程式碼，將就業分成兩大類，分別是將 1 (全職工作)、2 (兼職工作) 歸類在 A (有工作)有 102 人和 3 (退休)、4 (目前沒有工作)歸類在 B (沒有工作)有 594 人，也把身體健康分成兩類，1 (優)、2 (很好)、3 (良好)歸類在 X (身體健康優良)有 554 人;把 4 (一般)、5 (差)歸類在 Y (身體健康差)有 142 人，以此來分析健康狀況和各資料之間的關聯性。

因為睡眠困難這欄位資料在原始表格中出現的是:-1、1、2、3，但是在資料庫網頁上的欄位說明卻顯示 0、1 分別代表否、是，說明和欄位資料並不符合，所以我們根據 excel 統計出的資料得知 X 身體健康優良的人中，睡眠困難 1 有 33 位、睡眠困難 2 有 223 位、睡眠困難 3 有 298 位；而 Y 身體健康差的人中，睡眠困難 1 有 26 位、睡眠困難 2 有 62 位、睡眠困難 3 有 54 位，所以我們根據常理分析睡眠困難欄位的 1 代表的是有睡眠困難，2 代表的是普通，3 代表的是沒有睡眠困難。

<各欄位數字代表意思>

年齡	身體健康狀況	精神健康	牙科健康	就業	壓力是否影響患者的睡眠
患者年齡 1 : 50-64 2 : 65-80	病人身體健康自我評估 -1： 拒絕 1： 優 2： 很好 3： 良好 4： 一般 5： 差	病人精神或心理健康狀況的自我評價 -1： 拒絕 1： 優 2： 很好 3： 良好 4： 一般 5： 差	患者口腔或牙齒健康狀況的自我評估 -1：拒絕 1：優 2：很好 3：良好 4：一般 5：差 6：超差	病人的就業狀況或工作相關資訊 -1： 拒絕 1： 全職工作 2： 兼職工作 3： 退休 4： 目前沒有工作	0： 否 1： 是

藥物是否影響病人的睡眠	身體疼痛是否干擾患者睡眠	使用沐浴的需求是否影響病人的睡眠	不明因素是否影響患者睡眠	睡眠困難	處方_睡眠_藥物	性別
0：否 1：是	0：否 1：是	0：否 1：是	0：否 1：是	-1：拒絕 1：有 2：普通 3：沒有	有關為患者開立的任何睡眠藥物的資訊 -1：拒絕 1：定期使用 2：偶爾使用 3：不使用	患者性別認同 1：男 2：女

四、資料集的統計分析

(壹)entropy

我們先計算身體健康和每個欄位的聯合熵(entropy)來衡量他們之間的關聯性/不確定性。檢查是否有些變量可能與身體健康之間幾乎沒有任何關聯，或者幾乎沒有變化。如下:

Physical_Health 和 Age 的聯合熵為 0.5059291450508359

Physical_Health 和 Mental_Health 的聯合熵為 1.682371442459802

Physical_Health 和 Dental_Health 的聯合熵為 2.0492608514088757

Physical_Health 和 Employment 的聯合熵為 0.915764554760258

Physical_Health 和 Stress_Keeps_Patient_from_Sleeping 的聯合熵為 1.0662210656265967

Physical_Health 和 Medication_Keeps_Patient_from_Sleeping 的聯合熵為 0.7167137143651758

Physical_Health 和 Pain_Keeps_Patient_from_Sleeping 的聯合熵為 1.0080915709837188

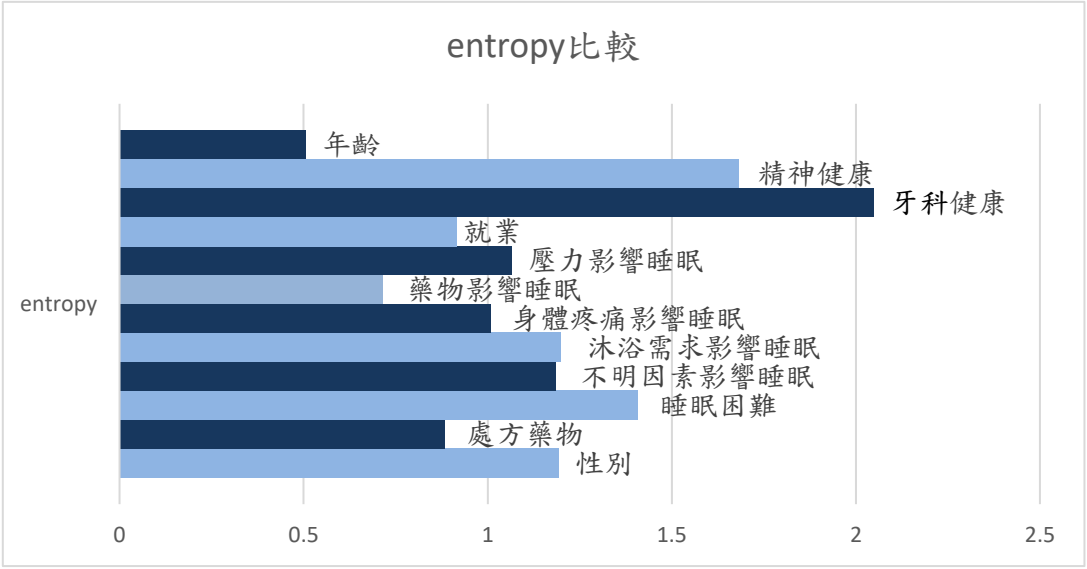
Physical_Health 和 Bathroom_Needs_Keeps_Patient_from_Sleeping 的聯合熵為 1.198888957183537

Physical_Health 和 Uknown_Keeps_Patient_from_Sleeping 的聯合熵為 1.18493403035435

Physical_Health 和 Trouble_Sleeping 的聯合熵為 1.4089762495320692

Physical_Health 和 Prescription_Sleep_Medication 的聯合熵為 0.8840484797685908

Physical_Health 和 Gender 的聯合熵為 1.1938990398684628



從計算出的 entropy 中相互比較後得出 entropy 最高和最低的變數分別是牙科健康跟年齡。Age 和 Physical_Health 的 entropy 為 0.5059，這表示他們之間的共同不確定性最低，也可能是因為我們資料中的年齡層都是在同一個區間才導致這個結果。而 Dental_Health 和 Physical_Health 的聯合熵為 2.0493，這表示他們之間的共同不確定性最高，關聯性也最弱，是此次分析中不確定性最高的變數，因此在之後的統計分析中牙科健康不列入考慮。

✧ 決策樹

我們使用上面計算出的 entropy 用 python 做一個預測身體健康狀況的決策樹。當我們輸入一筆想要預測的數據，可以使用程式碼得到身體健康狀況(1~5)的預測結果，模型準確度為 0.4784688995215311。

例如:

```
# 定義要進行預測的輸入數據
test_data = pd.DataFrame({
    'Age': [2],'Mental_Health': [3],'Dental_Health': [3],
    'Employment': [3],'Stress_Keeps_Patient_from_Sleeping': [0],
    'Medication_Keeps_Patient_from_Sleeping': [0],
    'Pain_Keeps_Patient_from_Sleeping': [0],
    'Bathroom_Needs_Keeps_Patient_from_Sleeping': [0],
    'Uknown_Keeps_Patient_from_Sleeping': [1],
    'Trouble_Sleeping': [2],'Prescription_Sleep_Medication': [3],
    'Gender': [2]})
```

得到: The predicted physical health is: 4

藉由計算 entropy 可得知其實每一欄位都跟身體健康有關聯，只有牙科健康較無關。那以下我們選取一些欄位和身體健康再做更仔細的百分比分析。

(貳)百分比

1. 身體狀況和精神狀況分析:

身體健康優良的 554 人中，精神狀況優的有 199 人佔 35.9206%、精神狀況很好的有 233 人佔 42.0578%、精神狀況良好的有 111 人佔 20.0361%、精神狀況一般的有 11 人佔 1.9856%、沒有精神狀況差的人。

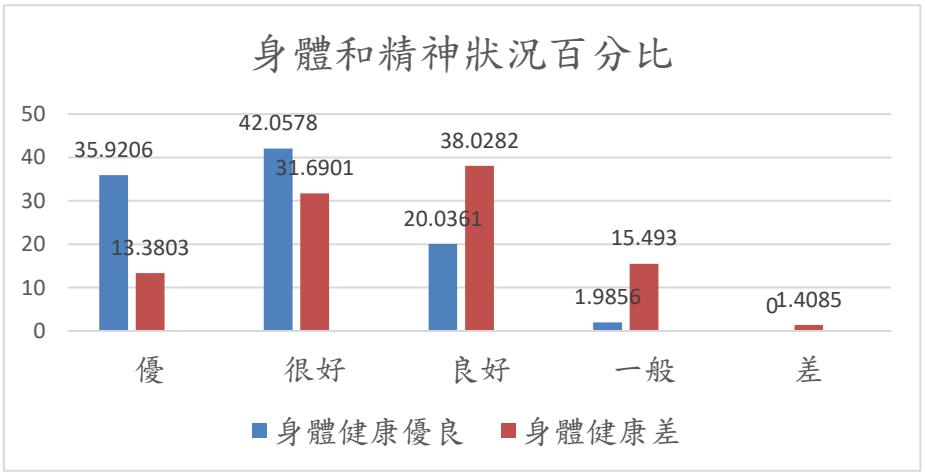
身體健康差的 142 人中，精神狀況優的有 19 人佔 13.3803%、精神狀況很好的有 45 人佔 31.6901%、精神狀況良好的有 54 人佔 38.0282%、精神狀況一般的有 22 人佔 15.4930%、精神狀況差的有 2 人佔 1.4085%。

由以下兩點:

(1)身體健康優的人中，精神狀況優的人佔了 35.9206%，和身體健康差的人中，精神狀況優的人佔 13.3803%來比較，身體健康優的人精神狀況也會比較好。

(2)可能是因為資料庫的數據量不夠龐大，所以會有些許落差，但是可以看到身體健康的人中，精神狀況優到差的人數大致上是一直在下降的，甚至沒有精神狀況差的人。

➜我們得知身體健康和精神狀況是有關連的。



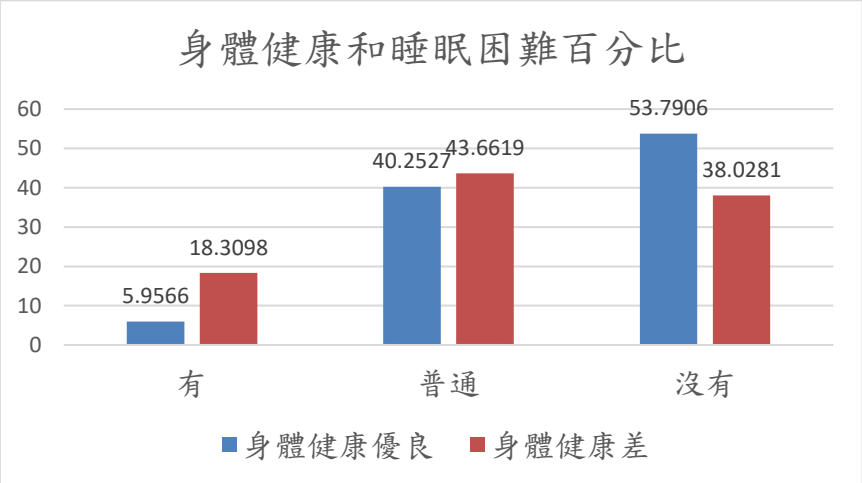
2. **身體狀況和睡眠困難分析：**

身體健康優良的 554 人中，有睡眠困難的有 33 人佔 5.9566%、睡眠困難普通的有 223 人佔 40.2527%、沒有睡眠困難的有 298 人佔 53.7906%。

身體健康差的 142 人中，有睡眠困難的有 26 人佔 18.3098%、睡眠困難普通的有 62 人佔 43.6619%、沒有睡眠困難的有 54 人佔 38.0281%。

在有睡眠困難的狀況下，18.3098% > 5.9566%，身體健康差的比例高於身體健康優良的比例；反之，在沒有睡眠困難的狀況下，53.7906% > 38.0281%，身體健康優良的比例高於身體健康差的比例。

➜ 我們得知身體健康和睡眠困難是有關連的。



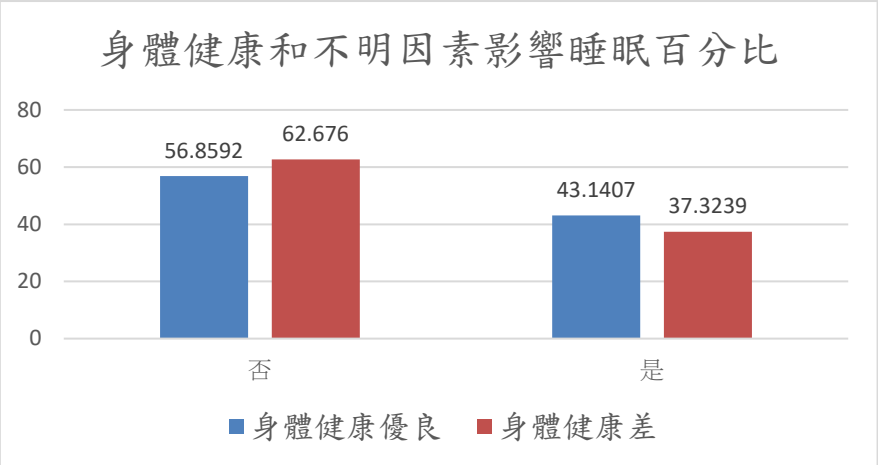
3. **身體狀況和不明因素影響睡眠分析：**

身體健康優良的 554 人中，沒有不明因素影響睡眠的有 315 人佔 56.8592%、有不明因素影響睡眠的有 239 人佔 43.1407%。

身體健康差的 142 人中，沒有不明因素影響睡眠的有 89 人佔 62.6760%、有不明因素影響睡眠的有 53 人佔 37.3239%。

從圖形上來看，比例沒有相差太多，所以我們認為不明因素影響睡眠和身體健康狀況並沒有太大的關聯，可能要再做更精細的不明原因分析，才能得到更有參考價值的數據。

➜我們得知身體健康和不明因素影響睡眠是沒有太大關聯的。



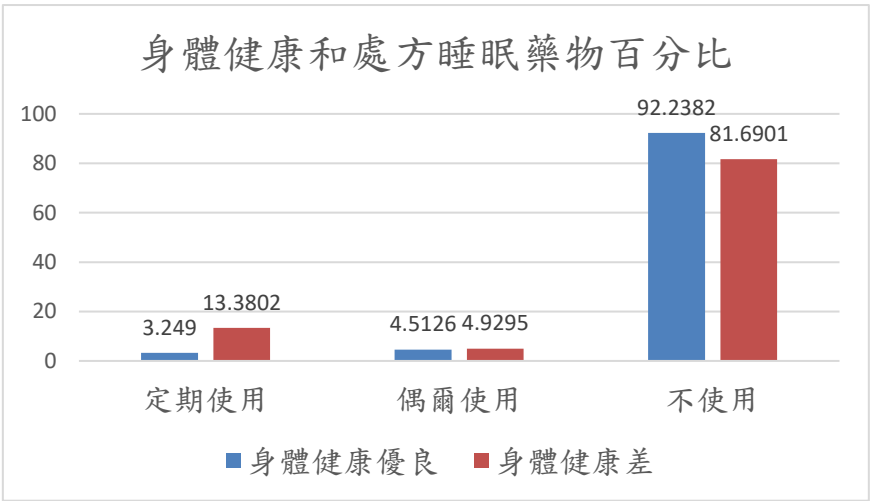
4. 身體狀況和處方睡眠藥物分析:

身體健康優良的 554 人中，有定期使用藥物的有 18 人佔 3.2490%、偶爾使用藥物的有 25 人佔 4.5126%、不使用藥物的有 511 人佔 92.2382%。

身體健康差的 142 人中，有定期使用藥物的有 19 人佔 13.3802%、偶爾使用藥物的有 7 人佔 4.9295%、不使用藥物的有 116 人佔 81.6901%。

從資料裡來看雖然絕大部分的人都沒有在使用藥物，但在定期使用藥物的人當中，身體健康差的人還是比身體健康優良的人比例上來得還要高(13.3% > 3.2%)。

➜得知雖然數據沒有差太多，但是身體健康的狀況多少還是和使用藥物有關係的。



(參) Odds Ratio

我們將參數視為兩個事件，評估某事件是否受到另一個事件影響，評估兩者之間的關聯強度· Odds Ratio > 1 時，代表在 A 發生的情況下 B 也容易發生。反之，當 Odds Ratio < 1 時，代表在 A 發生的情況下 B 不容易發生，而 Odds Ratio = 1 時，代表兩個之間無關。

用 EXCEL 統計出身體健康優之中壓力有影響睡眠的有 134 人，壓力沒有影響睡眠的有 39 人，身體健康差的人中壓力有影響睡眠的有 420 人，壓力沒有影響睡眠的有 103 人。用 python 畫出下列表格:

	身體差	身體優
受壓力影響睡眠	420	134
沒有受壓力影響睡眠	103	39

$$\text{Odds Ratio} = \frac{39/134}{103/420} = 1.186784$$

➜得知壓力有影響睡眠比壓力沒有影響睡眠容易身體健康差高 1.18 倍。

用 EXCEL 統計出身體健康優當中，受藥物影響睡眠的有 24 人，沒有受藥物影響睡眠的有 530 人。身體健康差的人當中，受藥物影響睡眠的有 15 人，沒有受藥物影響睡眠的有 127 人。用 python 畫出下列表格:

	身體差	身體優
有藥物影響睡眠	15	24
無藥物影響睡眠	127	530

$$\text{Odds Ratio} = \frac{15/24}{127/530} = 2.608268$$

➜得知藥物有影響睡眠比藥物沒有影響睡眠容易身體健康差高約 2.6 倍。

用 EXCEL 統計出身體健康優之中，身體疼痛有影響睡眠的有 95 人，身體疼痛沒有影響睡眠的有 459 人，身體健康差的人之中，身體疼痛有影響睡眠的有 57 人，身體疼痛沒有影響睡眠的有 85 人。

可以畫出下列表格:

	身體差	身體優
身體疼痛影響睡眠	57	95
身體疼痛沒有影響睡眠	85	459

$$\text{Odds Ratio} = \frac{57/95}{85/459} = 3.24$$

➜得知身體疼痛有影響睡眠比身體疼痛沒有影響睡眠容易身體健康差高約 3.24 倍。

用 EXCEL 統計出身體健康優之中，沐浴需求有影響睡眠的有 278 人，沐浴需求沒有影響睡眠的有 276 人，身體健康差的人之中，沐浴需求有影響睡眠的有 74 人，沐浴需求沒有影響睡眠的有 68 人。

可以畫出下列表格:

	身體差	身體優
沐浴需求影響睡眠	74	278
沒有沐浴需求影響睡眠	68	276

$$\text{Odds Ratio} = \frac{74/278}{68/276} = 1.08$$

➜得知沐浴需求有影響睡眠比沐浴需求沒有影響睡眠容易身體健康差高約 1.08 倍。

小結:

影響身體健康程度由大到小分別是:身體疼痛影響睡眠、受藥物影響睡眠、壓力影響睡眠、沐浴需求影響睡眠。身體疼痛有很多種原因，常見的除了身體勞累引發的疼痛外，還有炎症引發的局部不適、精神因素產生的精神性疼痛等。由上述可知把身體顧好不要讓產生身體疼痛、吃藥的時候看藥物的副作用是否會影響睡眠、不要給自己過多的壓力，除了可以擁有更好的睡眠品質外，還可以讓身體更健康一些。

五、總結

在這份資料庫分析報告中，我們深入探討了身體健康狀況受何影響的問題，並運用了 entropy、不同健康狀況在各欄位中的百分比和 odd ratio 等方法進行了詳細分析。通過這些方法，我們能夠知道不同因素對健康狀況的影響程度和相互關聯性。

此外，我們使用了 odd ratio 來評估不同因素對於特定健康狀況的相對風險，這有助於識別潛在的危險因素。這些結果為健康政策制定者、醫療專業人員和個人提供了參考價值，可以指導他們在促進健康和預防疾病方面往正確的方向前進。

綜合而言，我們的分析突顯了身體健康狀況是多個因素影響的結果，還需要綜合考慮其他因素，透過更深入的數據分析和統計，才可以更好地理解這些影響因素之間的複雜關係，為促進健康和提高生活質量提供更有效的策略和措施。

六、附件程式碼

前處理-1.py

```python
import pandas as pd

# 讀取CSV檔案
file_path = 'C:/Users/annie/OneDrive/Desktop/NPHA-doctor-visits.csv'
data = pd.read_csv(file_path, encoding='big5')

# 刪除指定欄位
columns_to_drop = ['Number of Doctors Visited', 'Age', 'Race']
data = data.drop(columns=columns_to_drop)

# 刪除含有-1的列
data_cleaned = data[(data != -1).all(axis=1)]

# 保存清理後的資料
cleaned_file_path = '數據報告可修改資料(刪除-1檔案).csv'
data_cleaned.to_csv(cleaned_file_path, index=False)

print(f"清理後已保存為: {cleaned_file_path}")
```

前處理完整版.py

```python
import pandas as pd

file_path = 'C:/Users/annie/OneDrive/Desktop/數據報告可修改資料.csv'
data = pd.read_csv(file_path, encoding='big5')

# 刪除包含-1的行
data_cleaned = data[(data != -1).all(axis=1)]

# 定義分類函數
def categorize_employment(status):
    if status in [1, 2]:
        return 'A'
    elif status in [3, 4]:
        return 'B'
    else:
        return 'unknown'

# 應用分類函數
data_cleaned['Employment'] = data_cleaned['Employment'].apply(categorize_employment)

# 修改Physical_Health列的值
def categorize_physical_health(health):
    if health in [1, 2, 3]:
        return 'X'
    elif health in [4, 5]:
        return 'Y'
    else:
        return health

data_cleaned['Physical_Health'] = data_cleaned['Physical_Health'].apply(categorize_physical_health)

# 保存清理和分類後的數據到新的CSV文件
output_file_path = 'C:/Users/annie/OneDrive/Desktop/處理後的數據報告.csv'
data_cleaned.to_csv(output_file_path, index=False, encoding='big5')

print(f"清理和分類後已保存為: {output_file_path}")
```

```python
import pandas as pd
from scipy.stats import entropy

# 加載數據
file_path = 'C:/Users/user/OneDrive/桌面/處理後的數據報告.csv'
data = pd.read_csv(file_path)

# 修正欄位名稱中的錯誤
data.columns = data.columns.str.replace('\t', '')

# 欄位名稱列表
columns = ['Age', 'Physical_Health', 'Mental_Health', 'Dental_Health', 'Employment',
           'Stress_Keeps_Patient_from_Sleeping', 'Medication_Keeps_Patient_from_Sleeping',
           'Pain_Keeps_Patient_from_Sleeping', 'Bathroom_Needs_Keeps_Patient_from_Sleeping',
           'Uknown_Keeps_Patient_from_Sleeping', 'Trouble_Sleeping', 'Prescription_Sleep_Medication',
           'Gender']
# 計算聯合熵
def calculate_joint_entropy(df, var1, var2):
    # 確認 var2 在 df 中存在
    if var2 in df.columns:
        # 計算聯合概率分佈
        joint_prob = df.groupby([var1, var2]).size() / len(df)
        return entropy(joint_prob)
    else:
        return float('NaN')

# 遍歷每個欄位並計算與 Physical_Health 的聯合熵
joint_entropies = {}
for column in columns:
    if column != 'Physical_Health':  # 確保不計算自身和 Physical_Health 之間的聯合熵
        joint_entropy = calculate_joint_entropy(data, 'Physical_Health', column)
        joint_entropies[column] = joint_entropy
        if not pd.isnull(joint_entropy):
            print(f"Physical_Health 和 {column} 的聯合熵為: {joint_entropy}")
        else:
            print(f"無法計算 Physical_Health 和 {column} 的聯合熵")

joint_entropies
```

健康狀況和其他狀況人數.py

```python
import pandas as pd

# 讀取已處理的CSV文件
file_path = 'C:/Users/annie/OneDrive/Desktop/處理後的數據報告.csv'
data = pd.read_csv(file_path, encoding='big5')

# 計算 Physical_Health 為 X 和 Y 時，每個 Mental_Health 值的數量
mental_health_counts = data.groupby(['Physical_Health', 'Mental_Health']).size().unstack(fill_value=0)

# 計算 Physical_Health 為 X 和 Y 時，每個 Dental_Health 值的數量
dental_health_counts = data.groupby(['Physical_Health', 'Dental_Health']).size().unstack(fill_value=0)

Trouble_Sleeping_counts = data.groupby(['Physical_Health', 'Trouble_Sleeping']).size().unstack(fill_value=0)

Uknown_Keeps_Patient_from_Sleeping_counts = data.groupby(['Physical_Health', 'Uknown_Keeps_Patient_from_Sleeping']).size().unstack(fill_value=0)

Prescription_Sleep_Medication_counts = data.groupby(['Physical_Health', 'Prescription_Sleep_Medication']).size().unstack(fill_value=0)
# 打印結果
print("Mental Health Counts:")
print(mental_health_counts)
print("\nDental Health Counts:")
print(dental_health_counts)
print("\nTrouble_Sleeping Counts:")
print(Trouble_Sleeping_counts)
print("\nUknown_Keeps_Patient_from_Sleeping Counts:")
print(Uknown_Keeps_Patient_from_Sleeping_counts)
print("Prescription_Sleep_Medication Counts:")
print(Prescription_Sleep_Medication_counts)
```

決策樹準確度.py - C:/Users/annie/OneDrive/Desktop/決策樹準確度.py (3.12.3)

File Edit Format Run Options Window Help

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split

file_path = 'C:/Users/annie/OneDrive/Desktop/數據報告可修改資料(刪除-1).csv'
df = pd.read_csv(file_path)

# 清理列名中的多餘字符
df.columns = df.columns.str.strip()

# 定義特徵和目標變量
X = df.drop('Physical_Health', axis=1)
y = df['Physical_Health']

# 將數據分為訓練集和測試集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=39)

# 創建決策樹分類器並訓練
clf = DecisionTreeClassifier(random_state=39)
clf.fit(X_train, y_train)

# 評估模型在測試集上的準確性
accuracy = clf.score(X_test, y_test)
print(f"Accuracy={accuracy}")

# 定義要進行預測的輸入數據
test_data = pd.DataFrame({
    'Age': [2],
    'Mental_Health': [2],
    'Dental_Health': [2],
    'Employment': [3],
    'Stress_Keeps_Patient_from_Sleeping': [0],
    'Medication_Keeps_Patient_from_Sleeping': [0],
    'Pain_Keeps_Patient_from_Sleeping': [0],
    'Bathroom_Needs_Keeps_Patient_from_Sleeping': [1],
    'Uknown_Keeps_Patient_from_Sleeping': [0],
    'Trouble_Sleeping': [3],
    'Prescription_Sleep_Medication': [3],
    'Gender': [1]
})
# 使用模型進行預測
prediction = clf.predict(test_data)

predicted_health = prediction[0]
print(f"The predicted physical health is: {predicted_health}")
```

```
========== RESTART: C:/Users/annie/OneDrive/Desktop/決策樹準確度.py ==========
======
Accuracy=0.4784688995215311
The predicted physical health is: 2
```

參考資料:

資料庫來源:https://archive.ics.uci.edu/dataset/936/national+poll+on+healthy+aging+(npha)

Entropy: 輕鬆了解 Entropy(熵): 分類模型中評估變數的好幫手 - 書寫觀點.tw (notebookpage1005.blogspot.com)