身體健康狀況受何影響

11011142曾鈺涵、11011209謝佳璇

資料集介紹

- 資料庫名稱為全國健康老化民意調查(NPHA)。
- 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。而該資料集中每列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資料集的這個子集選擇了14個與健康和睡眠相關的特徵來預測任務。

資料集介紹

- 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
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

<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	身體疼痛是否影響患者睡眠	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

分析重點

使用百分比分析:

比較不同身體健康狀況和以下四項個別的關係

1. 精神狀況
2. 不明因素影響睡眠 (不明因素有可能是癌症、焦慮等等)
3. 有無睡眠困難
4. 處方睡眠藥物

分析重點

使用Odd Ratio分析:

身體健康狀況和以下四項個別的關係

1. 是否受壓力影響睡眠
2. 是否受藥物影響睡眠
3. 是否受身體疼痛影響睡眠
4. 是否受沐浴需求影響睡眠

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。

資料前處理

- (B) 特徵工程、選擇、擷取

接著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。

資料前處理

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，我們將就業分成兩大類

- 1 (全職工作)、2 (兼職工作) ➔ A (有工作) 共102人
- 3 (退休)、4 (目前沒有工作) ➔ B (沒有工作) 共594人

再把身體健康分成兩類

- 1 (優)、2 (很好)、3 (良好) ➔ X (身體健康優良) 共554人
- 4 (一般)、5 (差) ➔ Y (身體健康差) 共142人

以此來分析健康狀況和各資料之間的關聯性。

資料前處理

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

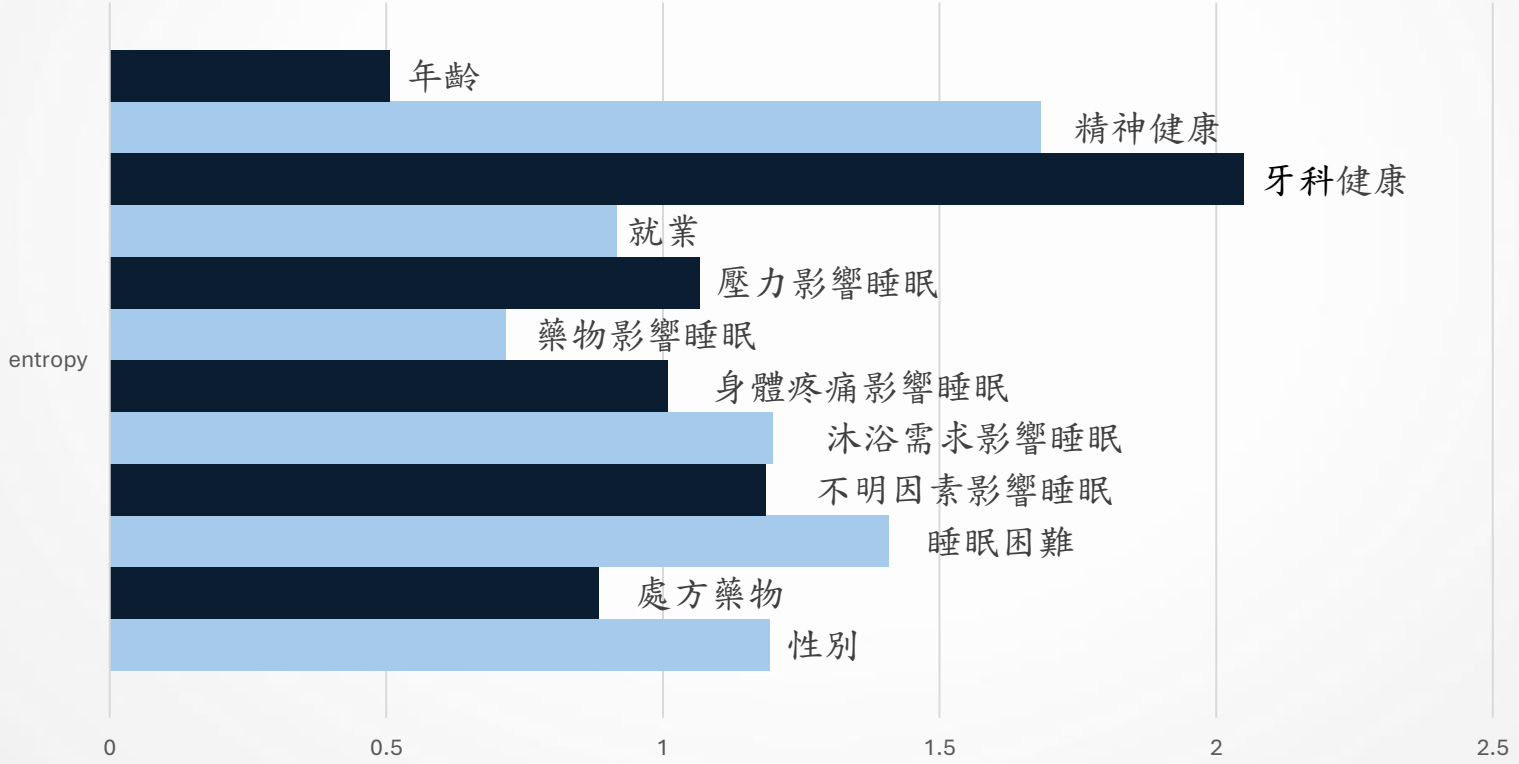
- X(身體健康優良)中，睡眠困難1有33位、睡眠困難2有223位、睡眠困難3有298位。
- Y(身體健康差)中，睡眠困難1有26位、睡眠困難2有62位、睡眠困難3有54位。

所以我們根據常理分析睡眠困難欄位:

- 1 ➔有睡眠困難
- 2 ➔普通
- 3 ➔沒有睡眠困難

資料集的統計分析(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。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:



資料集的統計分析(entropy)

entropy是計算不確定性的方法之一，用來衡量接收到新的資訊後，系統的雜亂程度是否會減輕，也就是系統能否更好的分類資料。entropy數值越大代表不確定性越大；反之， entropy數值越小代表不確定性越小。

為什麼會需要entropy呢？

建立模型可以提升模型效能，模型的準確度也會增加。

例如: entropy = 0 時，代表不確定性為 0，則可以清楚的分類出都是同一群；若是分類情況是 50% : 50%，則此時的不確定性是最大的。我們在製作決策樹時需要使用不確定性最小的變數，才可以製造出更有效率的決策樹。

資料集的統計分析(entropy)

- 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

  ➜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。
- Dental_Health和Physical_Health的entropy為2.0493

  ➜共同不確定性最高，關聯性最弱，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。

結論:

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。

資料集的統計分析(entropy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
```

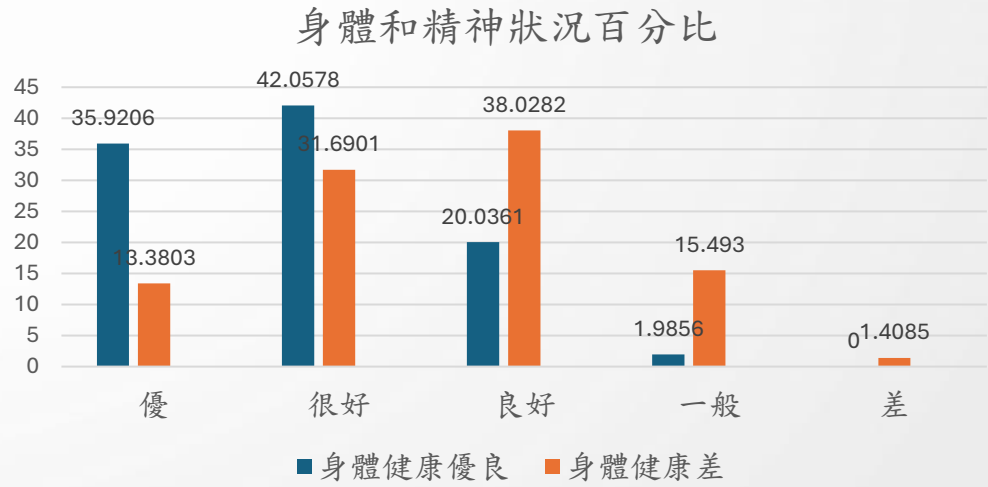
資料集的統計分析(百分比)

- **身體狀況和精神狀況分析**

(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。

➜由以上兩點可知身體健康和精神狀況是有關連的。



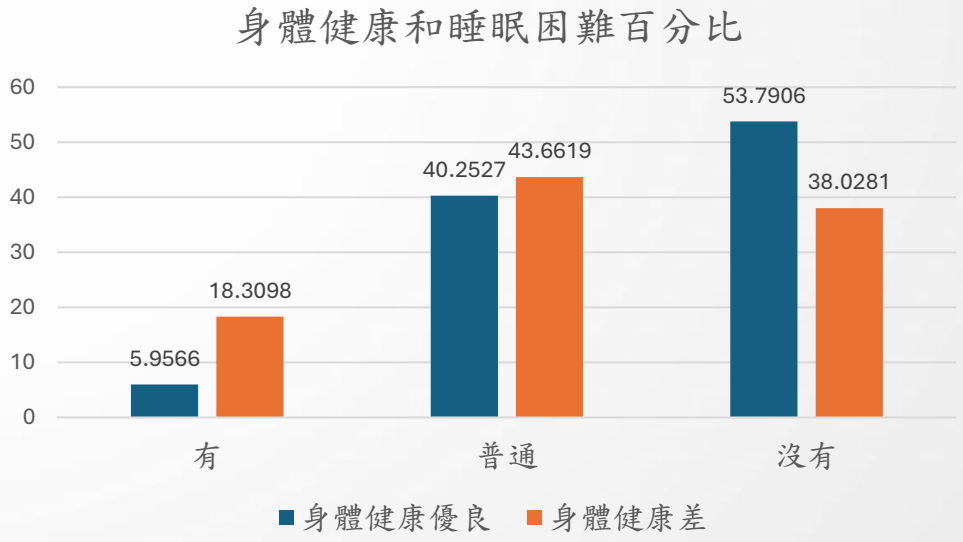
資料集的統計分析(百分比)

- **身體狀況和睡眠困難分析**

  在有睡眠困難的狀況下，身體健康差的比例高於身體健康優良的比例，18.31% > 5.96%。

  在沒有睡眠困難的狀況下，身體健康優良的比例高於身體健康差的比例，53.79% > 38.03%。

➜ 我們得知身體健康和睡眠困難是有關連的。

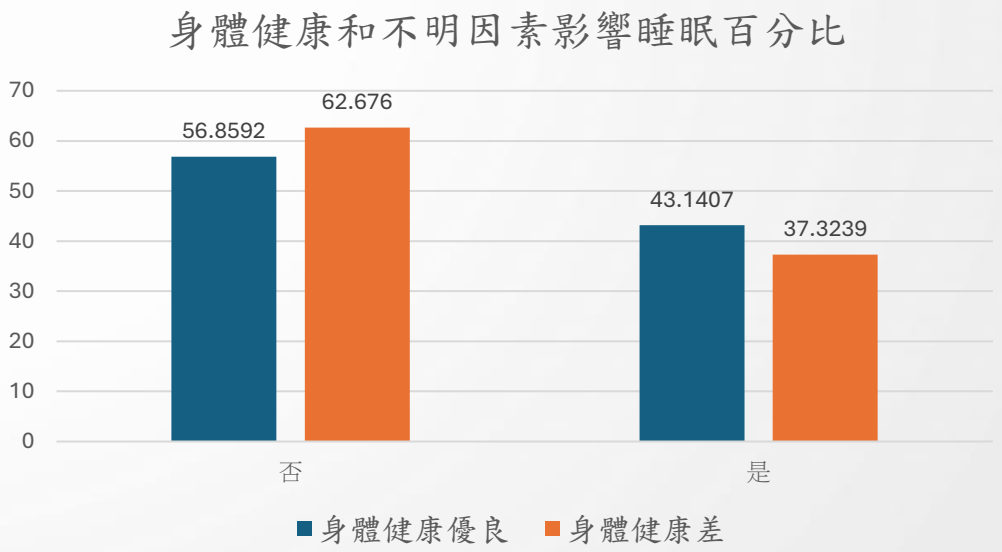


資料集的統計分析(百分比)

- **身體狀況和不明因素影響睡眠分析**

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。

➜得知身體健康和不明因素影響睡眠是沒有太大關聯的。

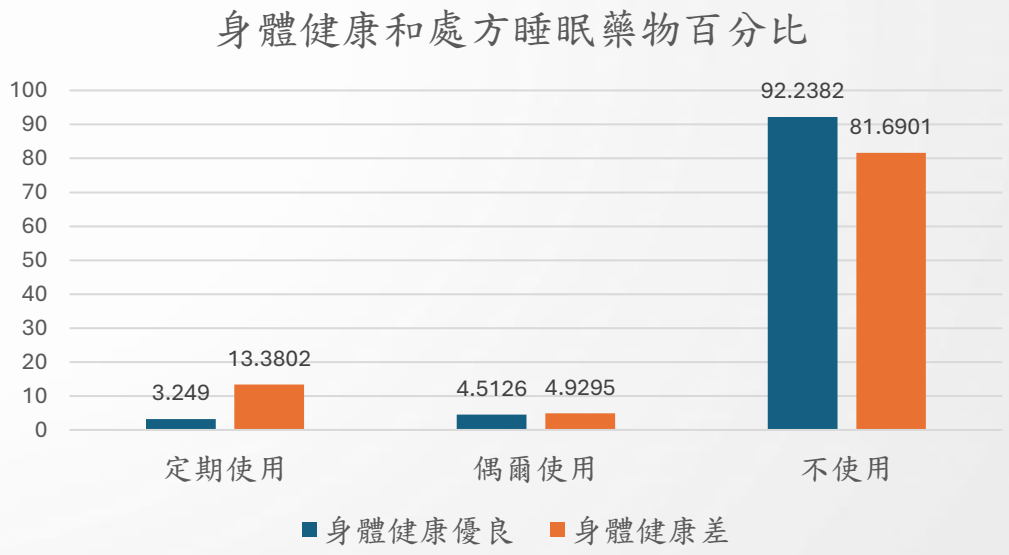


資料集的統計分析(百分比)

- **身體狀況和處方睡眠藥物分析**

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%)。

➜身體健康的狀況多少還是和使用藥物有關係的。



Odds Ratio

我們將參數視為兩個事件，評估某事件 A 是否受到另一個事件 B影響，評估兩者之間的關聯強度

- 當Odds Ratio > 1時，代表在A發生的情況下B也容易發生。
- 當Odds Ratio < 1時，代表在A發生的情況下B不容易發生。
- 當Odds Ratio = 1時，代表兩個之間無關。

資料集的統計分析(Odds Ratio)

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沒有影響睡眠容易身體健康差高1.18倍。

資料集的統計分析(Odds Ratio)

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沒有影響睡眠容易身體健康差高約2.6倍。

資料集的統計分析(Odds Ratio)

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。

用python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沒有影響睡眠容易身體健康差高約3.24倍。

資料集的統計分析(Odds Ratio)

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沒有影響睡眠容易身體健康差高約1.08倍。

Odd Ratio小結

影響身體健康程度 ⇨

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

(3.24 > 2.6 > 1.18 > 1.08)

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。

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。

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)


謝謝