**data mining報告**

**40747008S陳泓銪**

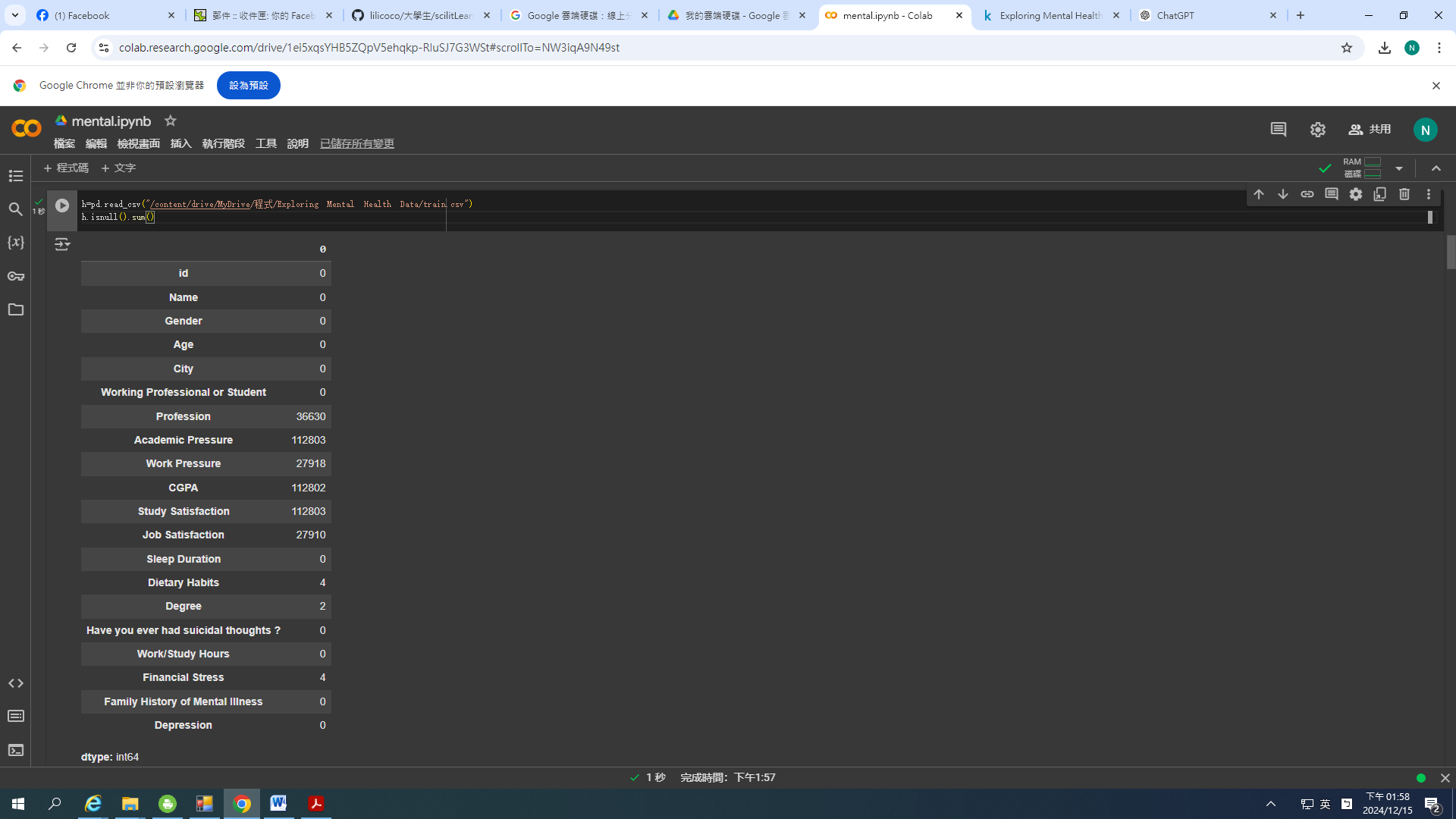
**1.前言**

在當代社會中，心理健康問題逐漸受到越來越多的關注，其中本專題需要預測的抑鬱症(Depression)，這是一種常見的精神健康障礙，已經對全球許多人的生活產生了深遠影響。

本資料集通過分析不同人之特徵如性別、年齡、城市、職業、睡眠時間、飲食習慣、工作或學習時長、經濟壓力、以及是否有家族精神疾病史等，建立機器學習模型預測抑鬱症的可能性。希望透過變量交互作用了解更多有關抑鬱症發病機理的線索，並幫助我們更好地理解哪些因素可能成為抑鬱症的風險因素。

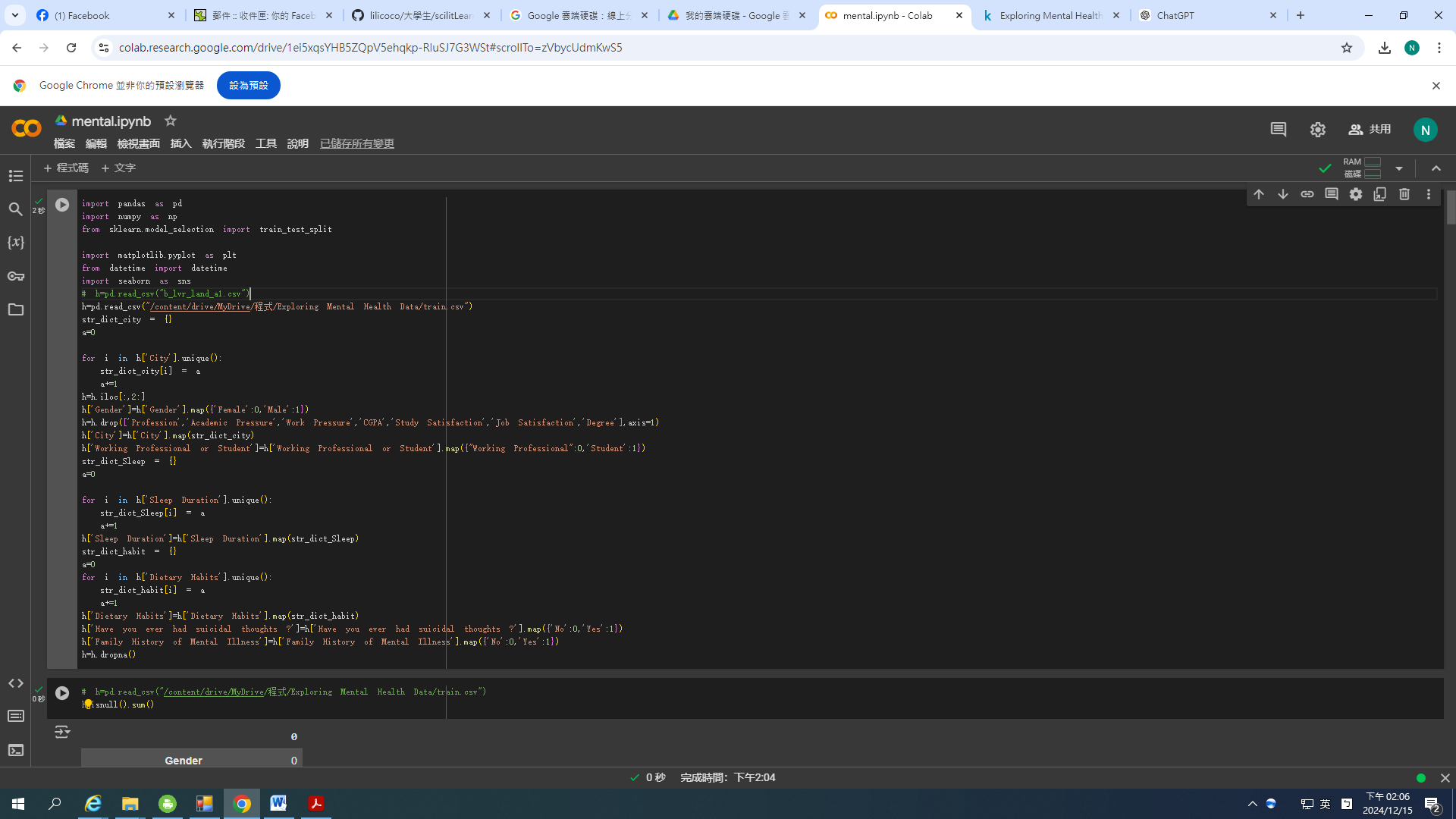
**2. 資料前處理**

該資料集由於是收集個人資訊，可能發生當事人不願意給部分資訊，導致這份資料集存在很多空值如下圖



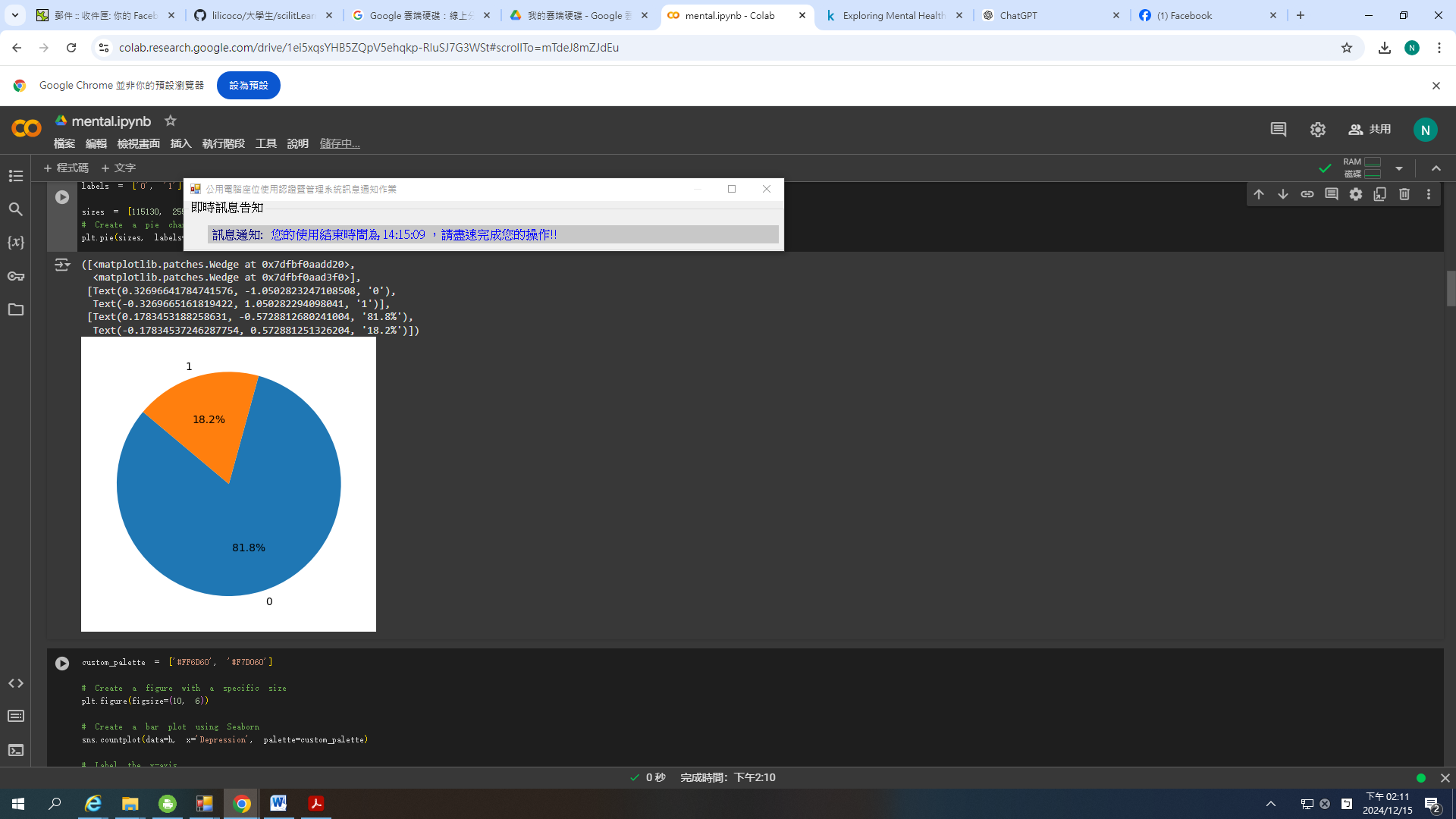
該資料集共有140700筆資料，其中Profession、Academic Pressure、Work Pressure、CGPA、Study Satisfaction、Job Satisfaction，以Academic Pressure為例子，細看其年紀分布，若為空值的資料通常年紀超過30以上，表示這問題不適用於出社會的人，而Work Pressure則不適用於太年輕的人，因此年輕族群才會很多空值，為了後續機器學習，本專題將這些存在大量空值的欄位直接刪除而不補值。

其他文字資料則轉換成數字參照下列程式碼

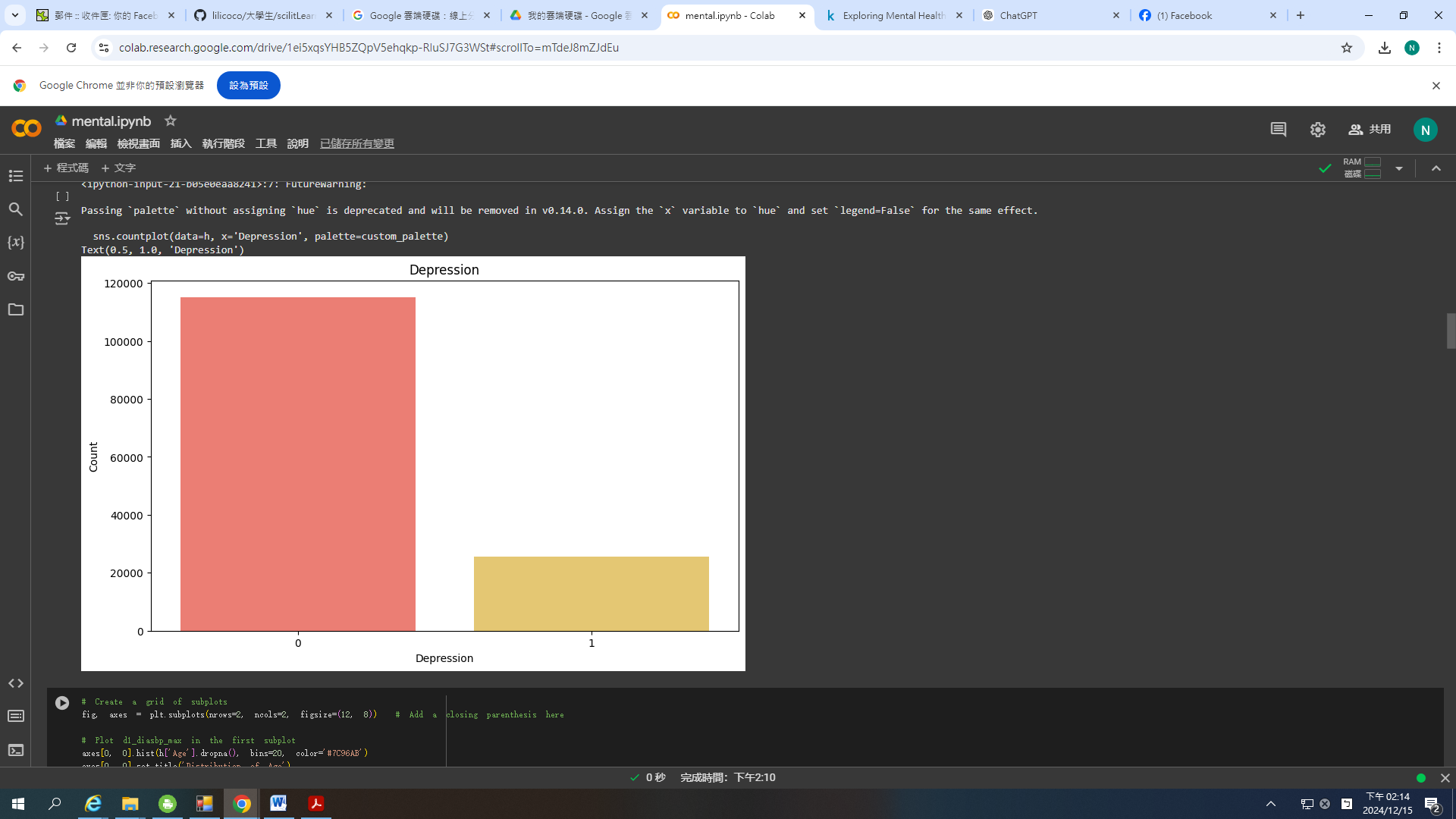


最終得到一份全部皆為數字且無空值的資料集

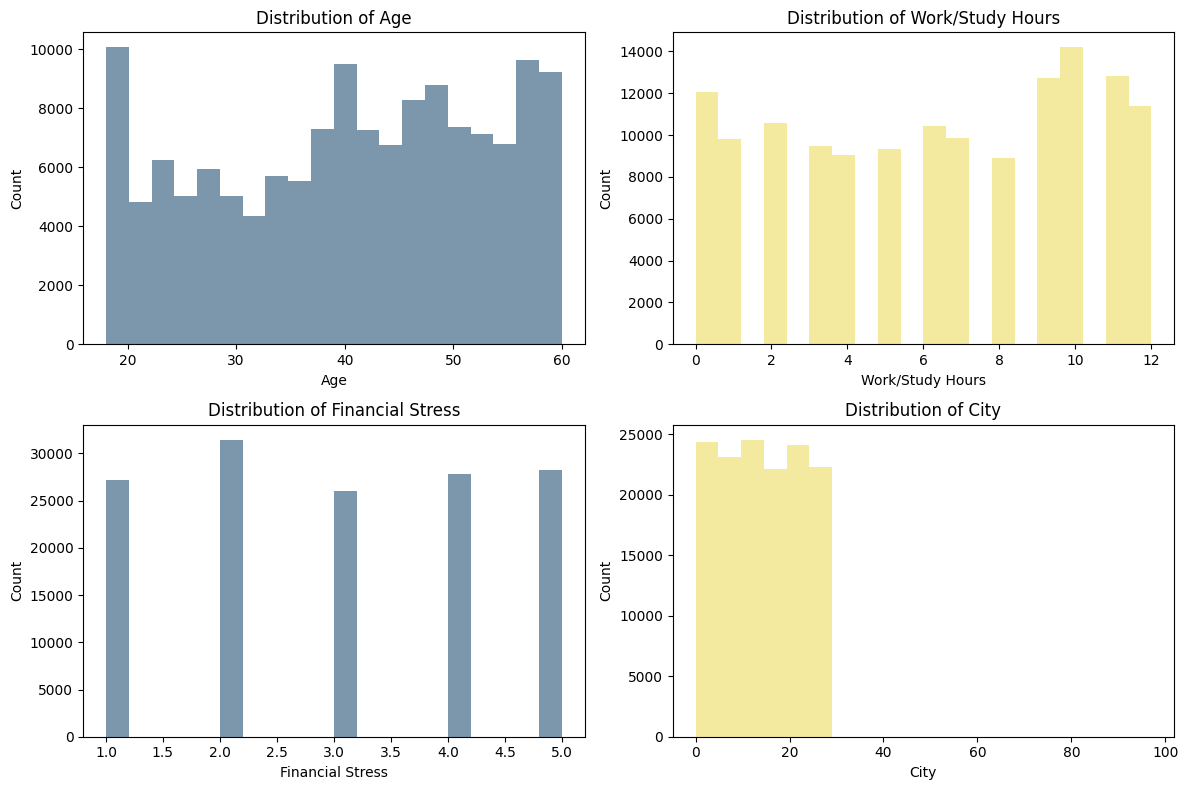
**3.資料可視化**



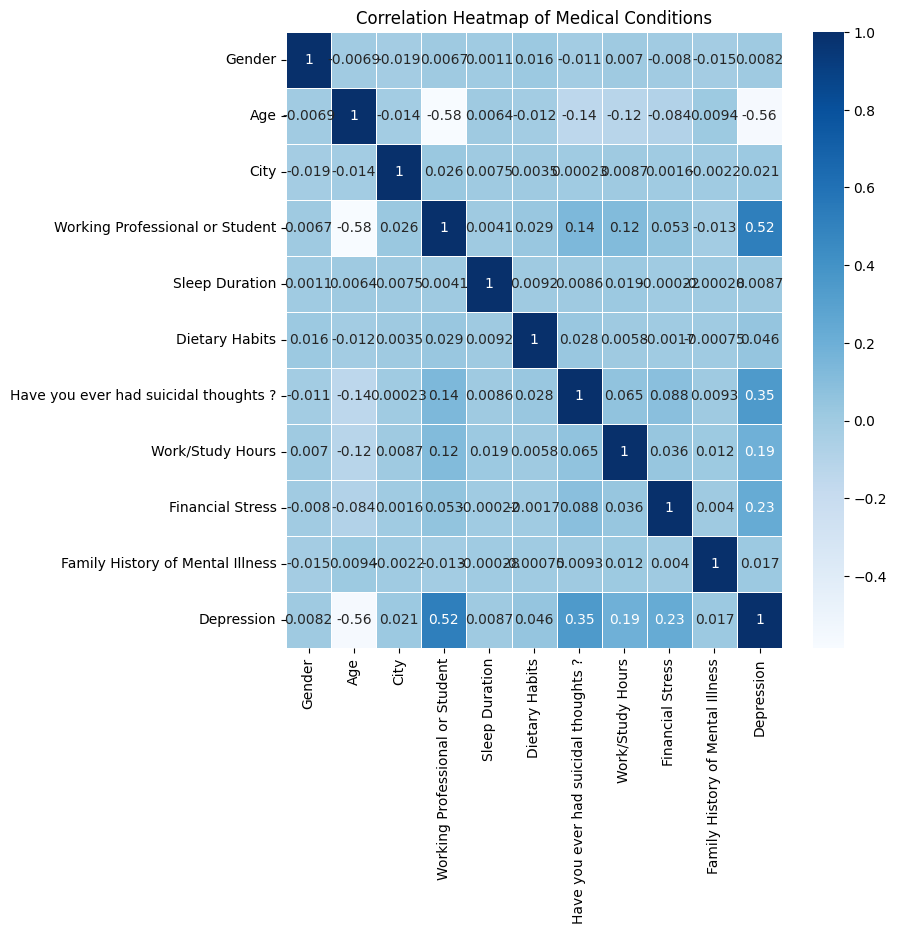
由圓餅圖可以得知資料不平衡，標註為1的資料過少，後面進行訓練時，會重複抽樣1的部分。



此圖顯示0與1數量差異大約快六倍。



此圖標示Age、Work/Study Hours、Financial Stress、City這幾個指標其數量分布，看起來相對平均，沒有某一種類別特別多。



最後我們將所有指標畫成correlation矩陣圖，觀察各個變量之關聯性，可以發現幾個比較有趣的點，例如Age與Depression呈現反相關，表示年齡越大越不容易有憂鬱症，反之年輕族群容易有憂鬱症，以20歲為例子有無憂鬱症有1330人，有憂鬱症為2184人，其他年紀如21-30歲都有這樣狀況。

### **1. RandomForestClassifier：**

* **呼叫方式**：由sklearn.ensemble導入RandomForestClassifier這個函式。
* **函式設定**：

|  |  |
| --- | --- |
| 參數 | 條件 |
| n\_estimators | 測試100及200 |
| max\_depth | 不設置及使用10 |

### **2. GradientBoostingClassifier：**

* **呼叫方式**：由sklearn.ensemble導入GradientBoostingClassifier這個函式。
* **函式設定**：

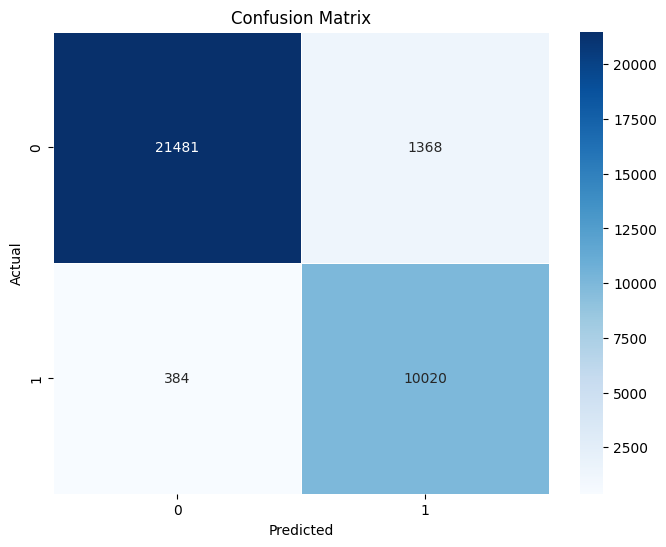
|  |  |
| --- | --- |
| 參數 | 條件 |
| n\_estimators | 測試100及200 |
| subsample | 測試1、0.9、0.5 |

**4.分類的效果評估**

**1. RandomForestClassifier：**

使用不同n\_estimators，100時準確率94%耗時18秒，200時準確率94%耗時29秒，由於樹木越多計算時間越長，但實際準確率沒有更高，因此採用100做下一步實驗，設定max\_depth為10，此時準確率為91%，由於max\_depth可以減少overfitting但也會降低準確率最終最佳參數選用以下

|  |  |
| --- | --- |
| 參數 | 條件 |
| n\_estimators | 100 |
| max\_depth | 不設置 |

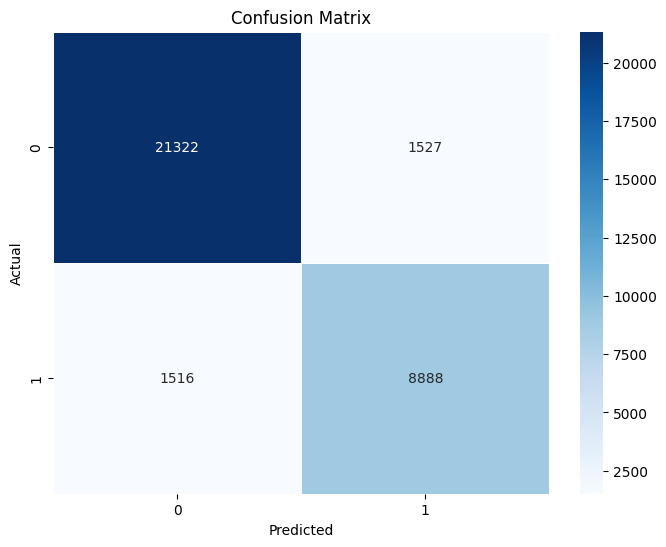


使用測試集的混淆矩陣，準確率為94.7%

**2. GradientBoostingClassifier：**

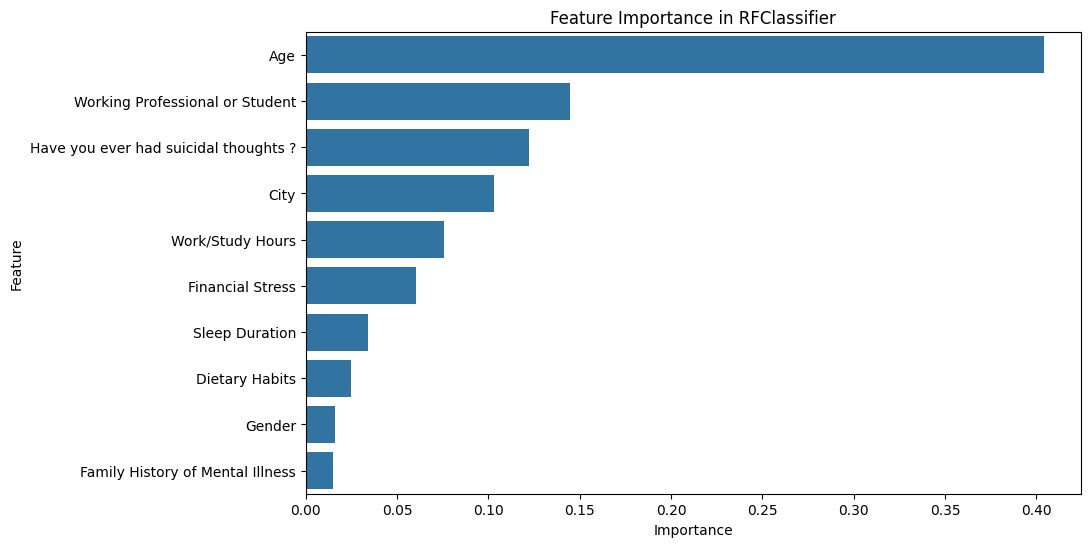
使用不同n\_estimators，100時準確率90%耗時7秒，200時準確率90%耗時26秒，由於樹木越多計算時間越長，但實際準確率沒有更高，因此採用100做下一步實驗，設定subsample為1、0.9、0.5，準確率皆為90%，最終最佳參數選用以下

|  |  |
| --- | --- |
| 參數 | 條件 |
| n\_estimators | 100 |
| subsample | 1 |



使用測試集的混淆矩陣，準確率為90%

使用RF 模型繪製重要性如下圖，表示年齡對於憂鬱症影響較大



**6. 結論**

本專題針對數據處理上，大部分採用有缺失就刪除的方式進行前處理，最終建立模型透過調整不同參數發現RandomForestClassifier準確率最好，而分析其重要性可以發現

Age對於分類是否憂鬱症影響最大