Python與機器學習 期末書面報告

遠距工作對心理健康的影響

M132040019 廖廣筑
M132040022 林士耕

大綱

一.摘要

這份研究報告主要在探討--遠距工作是否對於心理健康狀況有影響。本研究使用一個來自 Kaggle 的數據集，包含 5000 筆觀測值和 20 個變數，通過資料清洗、探索式資料分析（EDA）、特徵工程及多個機器學習模型，分析影響心理健康的關鍵因素。

本研究發現，每週工時、視訊會議次數與地區是心理健康預測中的關鍵特徵，進行多個模型比較，最後選擇 CatBoost 模型進行解釋與優化。

二.研究動機

隨著遠距工作逐漸成為現代職場的新常態，探討其對員工心理健康的影響更顯重要。我們希望透過分析相關數據，探索變數之間的關聯性，並進一步預測 Mental Health Condition（心理健康狀態），建立一個有效的預測模型，為未來公司對員工的健康管理提供參考價值，以推動更健康的職場文化。

三.資料介紹與研究方法

（一）資料集

- 資料來源：Kaggle，受訪者涵蓋全球員工。
- 含有 5000 筆資料與 20 項變數，每筆資料代表一位受訪者的相關紀錄。
- 涵蓋數值型變數與類別型變數，主要包括個人背景、工作環境、心理健康評估以及其他與工作生活平衡的相關指標。
- 主要特徵分類：

特徵分類	特徵欄位名稱
個人背景相關變數	年齡、地區、性別、壓力水平、睡眠品質、運動頻率、員工的唯一識別碼
工作相關變數	產業、每週工時、工作崗位、工作經驗、工作型態、生產力變化、視訊會議次數、遠端工作滿意度、公司對遠端工作的支持程度
心理健康相關變數	心理健康狀況、社會孤立評分、工作與生活平衡評級、是否獲取心理健康資源

（二）研究方法

- 目標變數：心理健康狀況（Mental_Health_Condition）
- 資料處理與分析：針對缺失值進行處理，使用 Label Encoding 與 Target Encoding 等技術進行數據清理及特徵工程。
- 實驗步驟與設定：

① EDA 分析數據分佈與目標變數之間的關聯性。

② 使用 SMOTE 方法解決資料不平衡問題。

③ 用 PyCaret 篩選模型。

④ 訓練多種模型，包括 CatBoost、LightGBM、Gradient Boosting 等，並比較其效能。

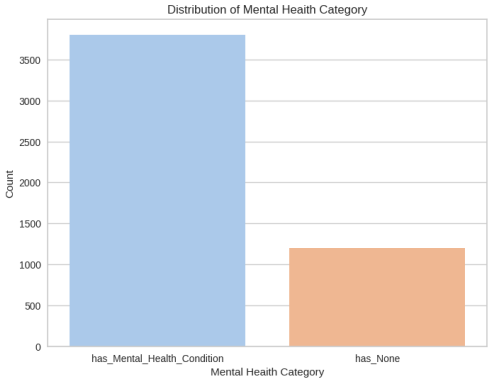
四.探索式資料分析

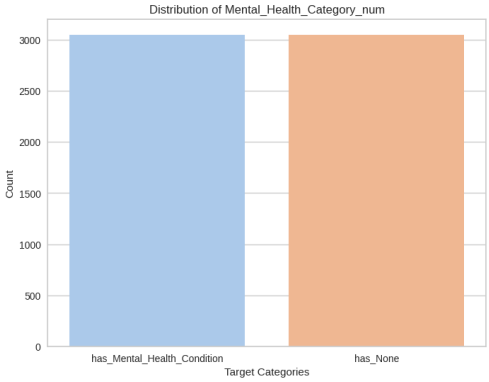
① 缺失值處理

- Mental_Health_Condition：此特徵形式有[焦慮、倦怠、憂鬱、nan]，這個的變數的 nan 表示沒有其他三種心理狀態，所以我們將 nan 填入 None。
- Physical_Activity：此特徵形式有[每日、每週、nan]，我們將 nan 全部填入 0，當成該員工沒有運動習慣。

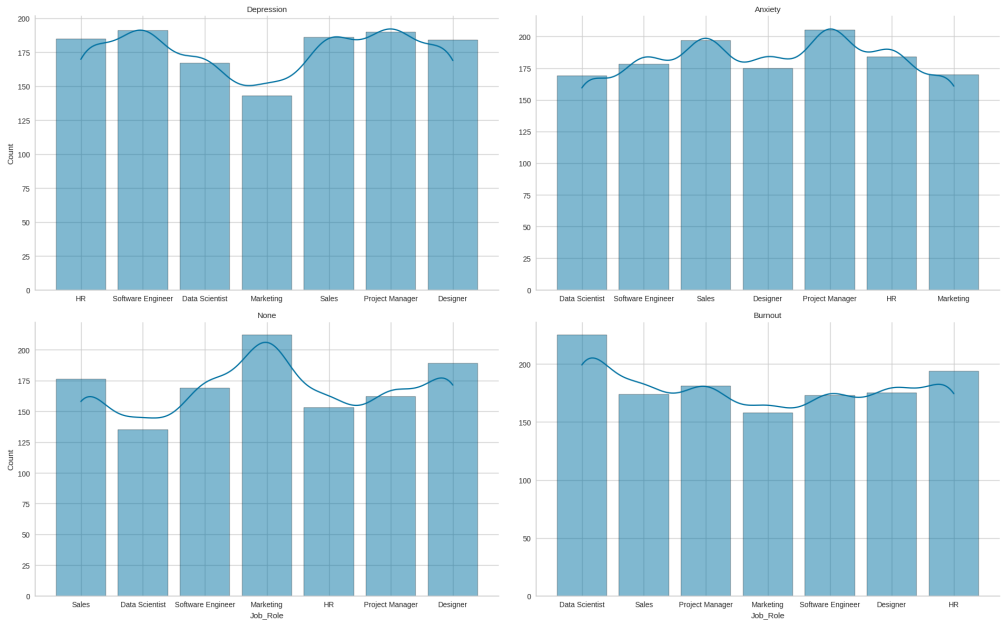
② 目標變數：**Mental_Health_Condition**

- 原始分佈（左下圖）：目標變數資料不平衡，有心理狀況的人數多於沒有心理狀況的人數。
- SMOTE 後目標變數分佈（右下圖）：目標變數分佈平均。





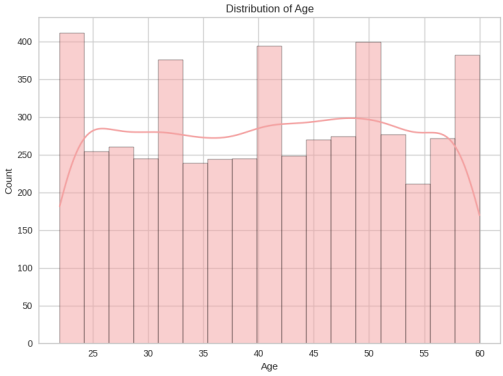
❖ 心理健康狀況 vs 職位

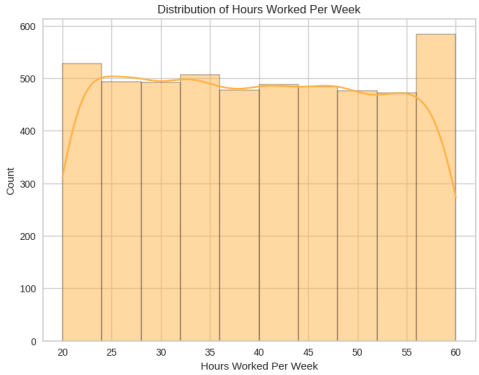


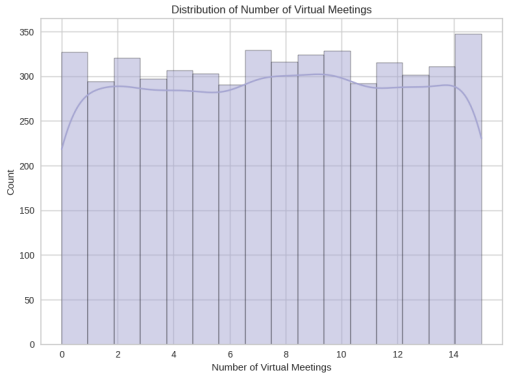
- Depression(憂鬱) 類別的長條圖（左上）中，軟體工程師和 PM 人數為最多，Sales 為最少。
- Anxiety(焦慮) 類別的長條圖（右上）中，PM 和 Sales 人數為最多。
- Burnout(倦怠) 類別的長條圖（右下）中，資料科學家人數為最多，行銷人員為最少。
- None(無以上三種心理狀況) 類別的長條圖（左下）中，行銷人員為最多，資料科學家為最少。

③ 數值型變數

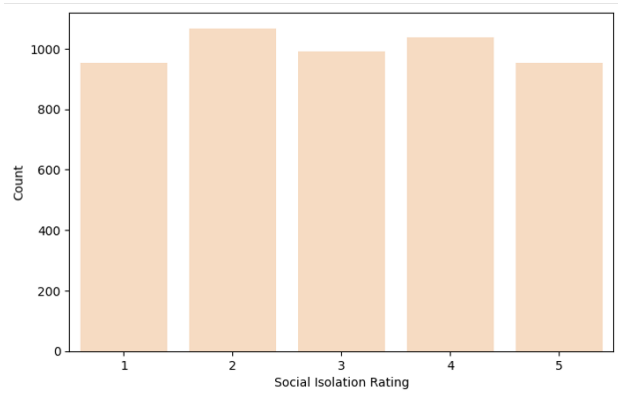
- 年齡分佈（左下圖）：分佈得很平均。
- 每週工時分佈（中間圖）：每週平均工時約 40 小時，亦分佈得很平均。
- 每週視訊會議次數分佈（右下圖）：分佈得很平均。

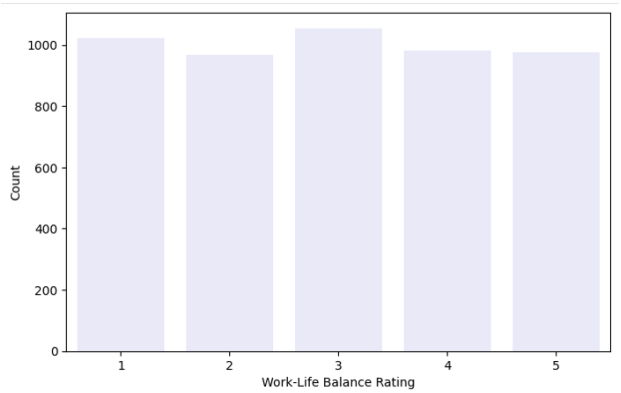




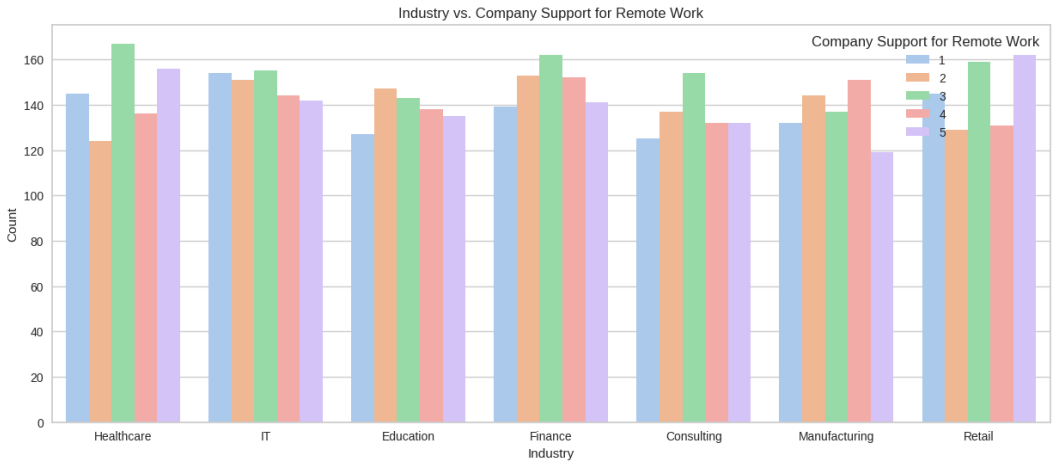


- 社會孤立指數（左下圖）：社會孤立指數出現最多次的是 2，第二多的是 4。
- 工作與生活平衡評級（右下圖）：Work_Life_Balance_Rating 代表的是工作與生活平衡評級，從此圖中可以看出 3 是最高的，代表大家對於這個平衡的感覺是處於中間；第二多的是 1，所以這分資料的員工沒辦法在工作與生活中取得較好的平衡。



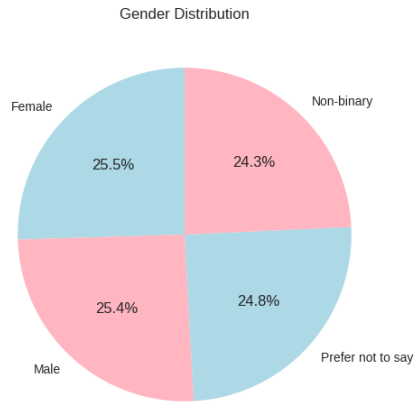


- 支持遠端工作指數 vs 產業
  - 支持遠端工作指數：數字越高代表公司越支持遠端工作。
  - 我們發現 3 是次數最多的，再來是 5，可以看出公司對遠端工作大多都是支持的立場。

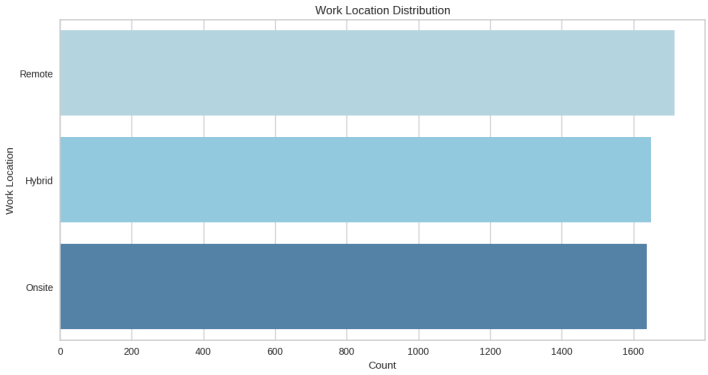


④ 類別型變數

- 性別：本資料集的性別變數有四種，分別為：女性、男性、非二元性別、不願透露。

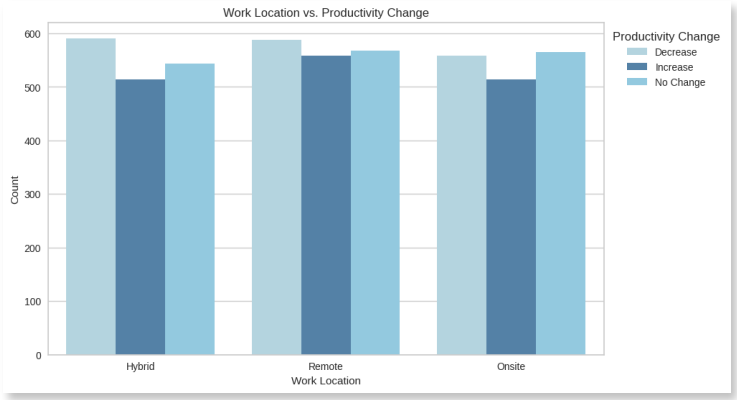


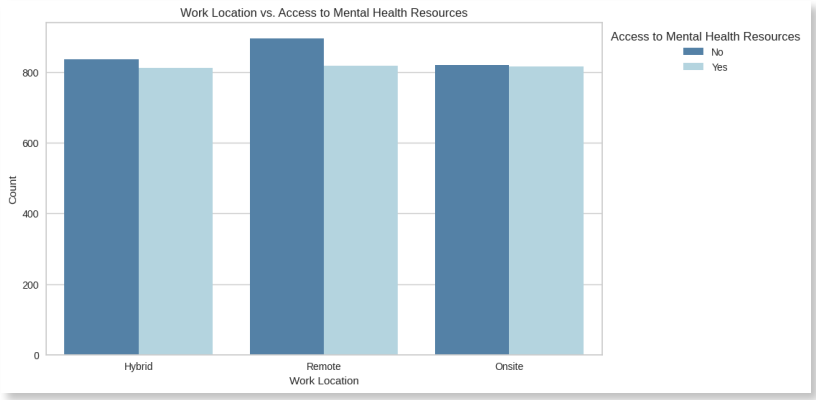
- 工作型態分佈：這份資料中，遠距工作總人數最多。



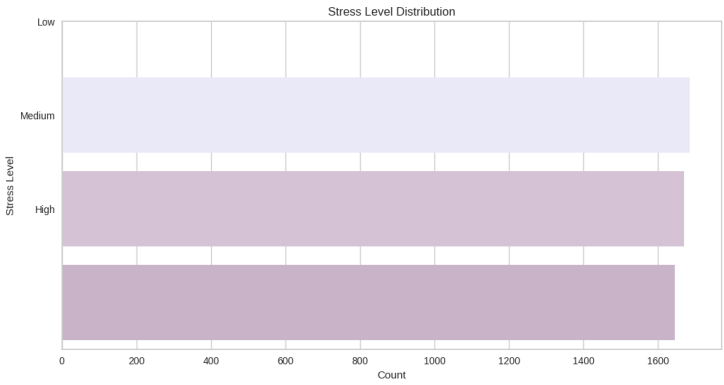
❖ 工作型態 vs 生產力改變（左下圖）：在遠距工作中，生產力增加跟減少的人數都很多。

❖ 工作型態 vs 是否獲取心理健康資源（右下圖）：遠距工作最少人去心理諮商。



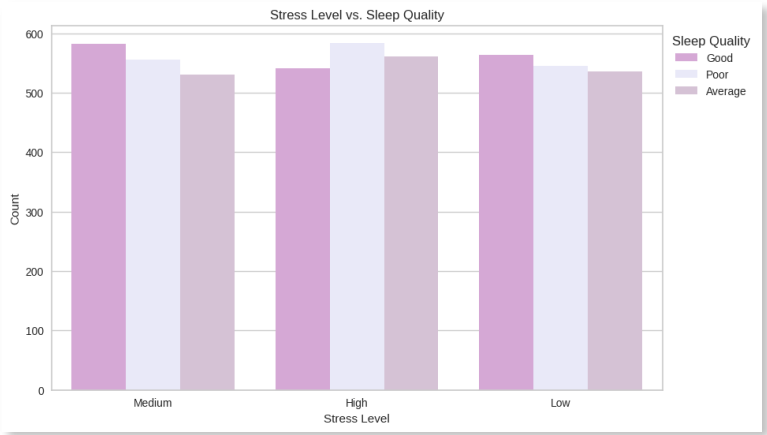


• 壓力指數分佈：

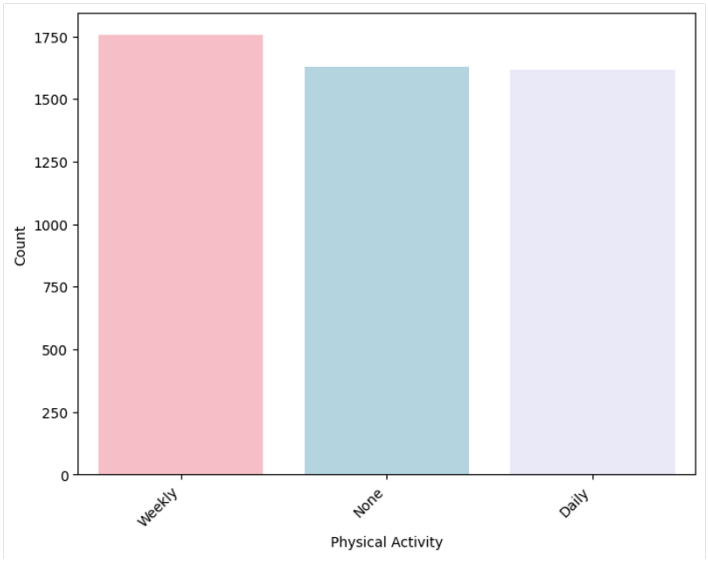


• 壓力 vs 生產力（左下圖）：在低壓的狀態下，最多人感覺會降低生產力而且最少人感覺會增加生產力。如果看增加生產力的值，可以發現在高壓的狀態下最多人會感覺生產力增加。

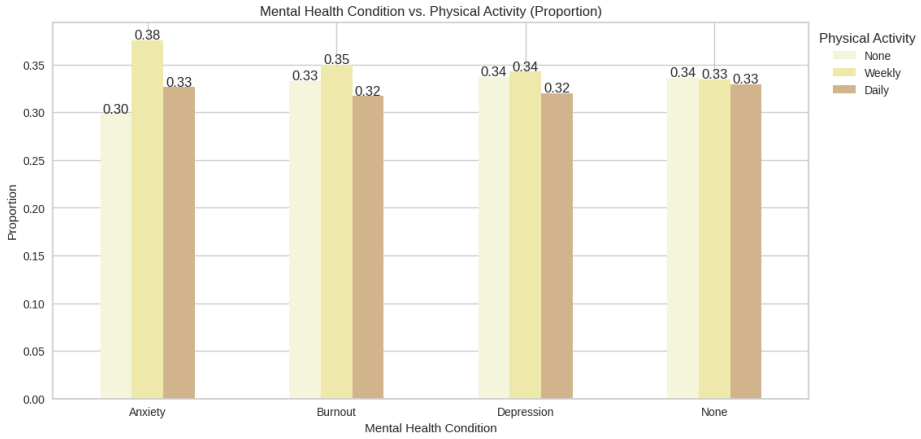
• 壓力 vs 睡眠品質（右下圖）：睡眠品質好的人中，最多人是壓力適中的，再來是低壓力。而睡眠品質不好的人中，最多人是處於高壓狀態。



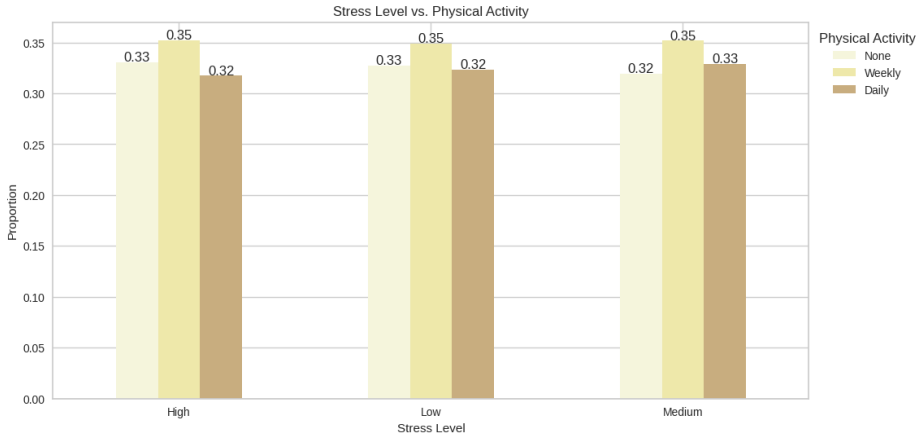
- 運動頻率分佈



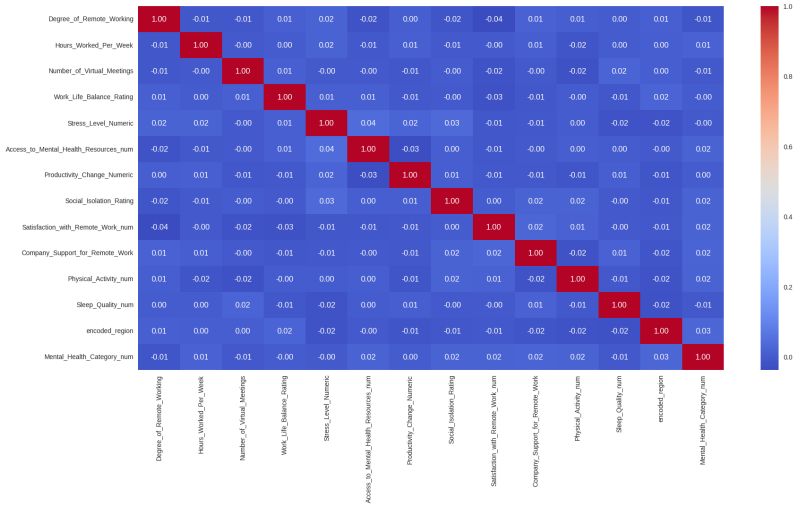
❖ 運動頻率 vs 心理健康狀態（比例）：在焦慮的人群中，每週運動習慣的比例是最高的。且每週運動的比例在三種負面情緒中的佔比都是最高的。



❖ 運動頻率 vs 壓力（比例）：每週運動頻率中，壓力指數低、中、高的比例相當。



⑤ 相關係數熱力圖



- 想法：熱力圖顯示所有變數之間的（線性）相關性都很低，表示這些變數之間可能沒有很強的線性關係，所以後續選擇模型的方向我們會優先選擇非線性模型。

五.特徵工程

- **Label Encoding**

變數名稱	處理前	處理後
Work_Location	Onsite/Hybrid/Remote	0 / 1 / 2
Stress_Level	Low/Medium/High	-1 / 0 / 1
Access_to_Mental_Health_Resources	Yes/No	1 / 0
Productivity_Change	Decrease/No Change/Increase	-1 / 0 / 1
Satisfaction_with_Remote _Work	Unsatisfied/Neutral /Satisfied	-1 / 0 / 1
Physical_Activity	None/Daily/Weekly	0 / 1 / 2
Sleep_Quality	Poor/Average/Good	-1 / 0 / 1
Mental_Health_Condition	has_Mental_Health _Condition/ has_None	1/0

- **Target Encoding**

變數名稱	處理前	處理後
Region	Europe / Asia / North America / South America / Oceania / Africa	0.7511627906976744, 0.7520184544405998, 0.7567567567567568, 0.7593712212817413, 0.7607142857142857, 0.7852834740651388.

六.模型建構與訓練

1. 模型建立步驟

⑥ 資料拆分：將數據拆分為 80% 訓練集、20% 測試集，random_state = 2024。

⑦ 數據平衡：使用 SMOTE 方法對訓練集進行數據平衡。

⑧ 建立模型：使用 PyCaret 中的分類器，從訓練集拆分出 20% 驗證集，並將數據標準化。

2. PyCaret 模型選擇

我們選擇的模型是 CatBoost、LightGBM、Gradient Boosting Classifier 和 Extreme Gradient Boosting，訓練後通過多項指標進行模型比較。

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.8136	0.8604	0.9152	0.7608	0.8308	0.6273	0.6407	4.1200
lightgbm	Light Gradient Boosting Machine	0.8089	0.8569	0.9058	0.7589	0.8258	0.6178	0.6299	0.2960
gbc	Gradient Boosting Classifier	0.8061	0.8461	0.9423	0.7409	0.8294	0.6121	0.6363	0.4790
xgboost	Extreme Gradient Boosting	0.8036	0.8624	0.8587	0.7736	0.8138	0.6072	0.6114	0.1460
rf	Random Forest Classifier	0.8001	0.8640	0.8514	0.7724	0.8098	0.6002	0.6037	0.9750
et	Extra Trees Classifier	0.7825	0.8520	0.8153	0.7653	0.7894	0.5650	0.5665	0.8310
ada	Ada Boost Classifier	0.7647	0.8307	0.8067	0.7443	0.7741	0.5294	0.5315	0.2260
knn	K Neighbors Classifier	0.7125	0.7963	0.5762	0.7925	0.6669	0.4250	0.4419	0.1870

由驗證集上的 Accuracy 準確度結果，發現 CatBoost 的各項數值都表現良好，LightGBM 次之，未來可以根據特定應用場景進一步優化這兩個模型。

3. 模型評估指標

- 各模型在測試集上的表現：

模型	Accuracy	AUC	Recall	Prec.	F1
CatBoost	0.7120	0.4810	0.9255	0.7500	0.8286
LightGBM	0.7270	0.4906	0.9388	0.7567	0.8380
Gradient Boosting Classifier	0.7230	0.4736	0.9481	0.7497	0.8373
Extreme Gradient Boosting	0.6890	0.4712	0.8723	0.7532	0.8084

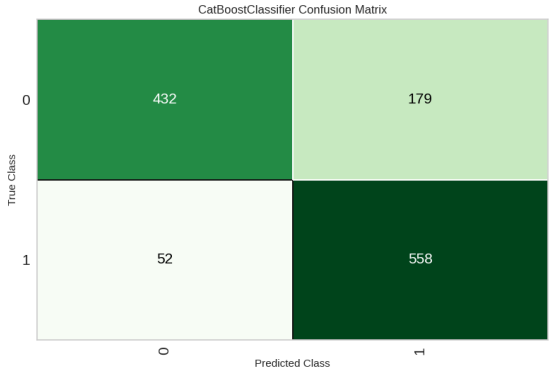
- 由上表可以發現，這些模型雖然 Recall 很高，但 AUC 值很低，表示模型的區分能力不佳。

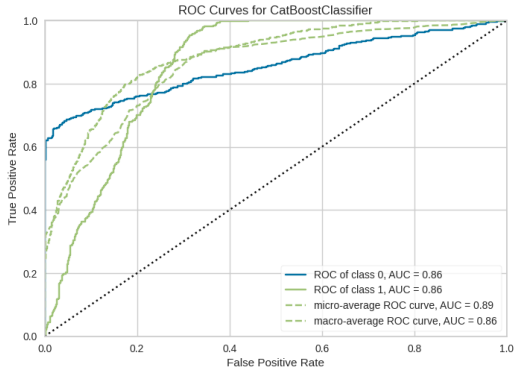
4. 模型簡介與比較

① CatBoost Classifier

✱ 簡介：CatBoost Classifier 是一種基於「梯度提升決策樹」的高效機器學習模型，特別適合結構化數據的分類問題，其自動化特徵處理和防止過擬合的能力佳，使其更廣泛應用在實務上。

✱ 混淆矩陣 &ROC Curves：



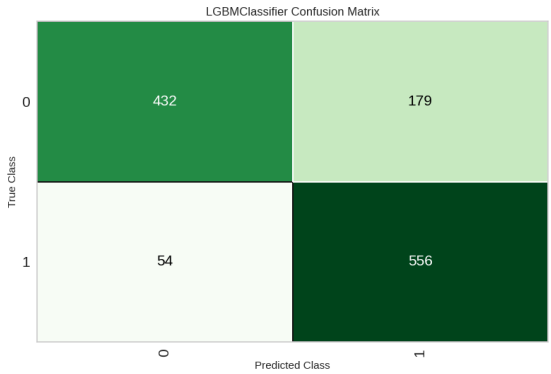


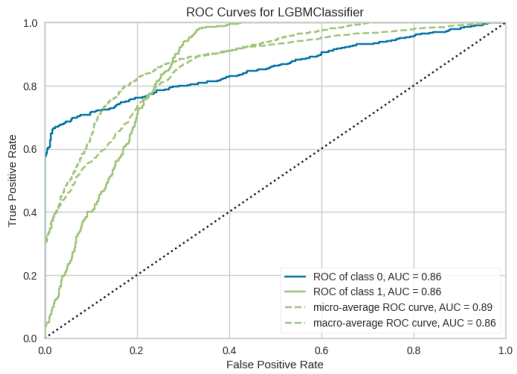
- 模型在預測類別 1 時表現較好，Recall 值達 **0.9255**，能有效捕捉正樣本。
- 假陽性數量較多，預測錯誤類別數占總預測類別的 **29.3%**。
- 整體 AUC 為 **0.86**，分類能力穩定。

② LightGBM

✱ 簡介：LightGBM 是一種基於「梯度提升框架」的高效機器學習模型，專為大規模數據集和高速計算設計，具有速度快、內存使用低、支持類別型特徵和並行處理等優勢。

✱ 混淆矩陣 &ROC Curves：



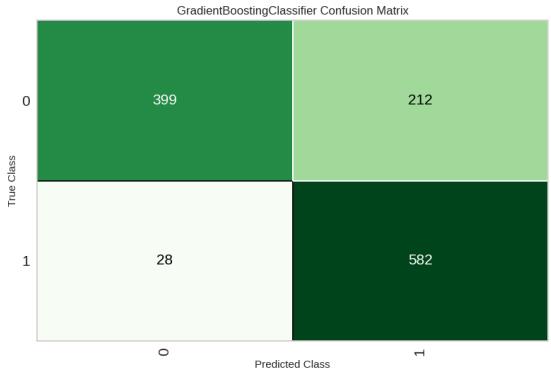


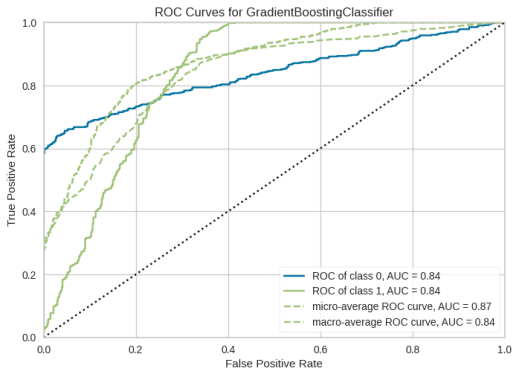
- 模型在預測類別 1 時表現不錯，Recall 值達 **0.9388**，捕捉正樣本能力強。
- 結果與 CatBoost 很相似，假陽性數量占總預測類別的 **29.3%**。
- 整體 AUC 為 **0.86**，與 CatBoost 表現相近。

③ Gradient Boosting Classifier

✱ 簡介：Gradient Boosting Classifier 是一種基於「梯度提升樹」的機器學習模型，通過逐步結合多個弱學習器（如決策樹）來提升整體模型性能，適合處理非線性數據和結構化數據的分類任務。

✱ 混淆矩陣 &ROC Curves：



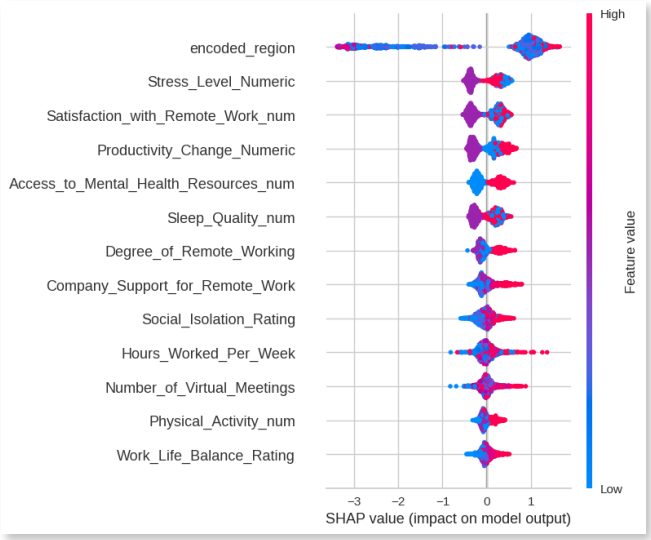


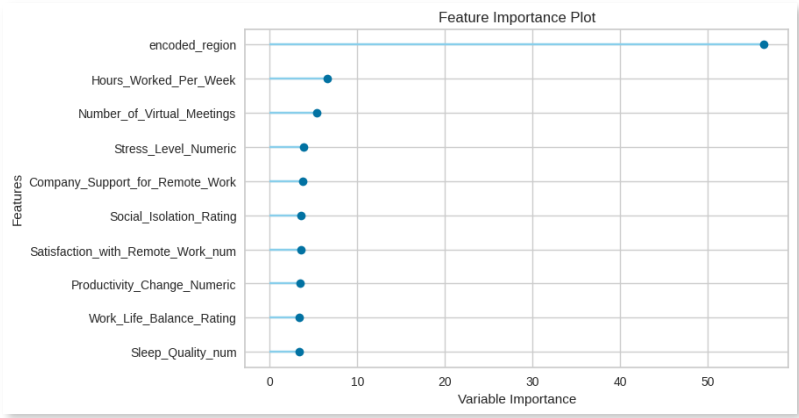
- 模型在預測類別 1 時 Recall 值達 **0.9481**，捕捉正樣本的能力最佳。
- 假陽性數量占總預測類別的 **34.7%**，錯誤率相對較高。
- 整體 AUC 為 **0.84**，區分能力略低於 CatBoost 和 LightGBM。

❖ **模型比較總結：**在多個模型中，CatBoost Classifier 在準確率、穩定性和 AUC 曲線上均表現最佳，適合作為首選模型；LightGBM 是次佳選擇。而 Gradient Boosting 更適合需要最大限度捕捉正樣本的情況下選擇。

5. 模型解釋性

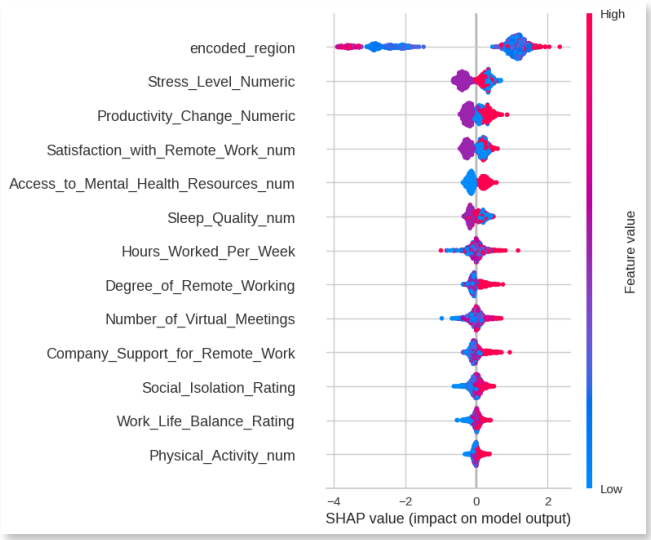
① CatBoost

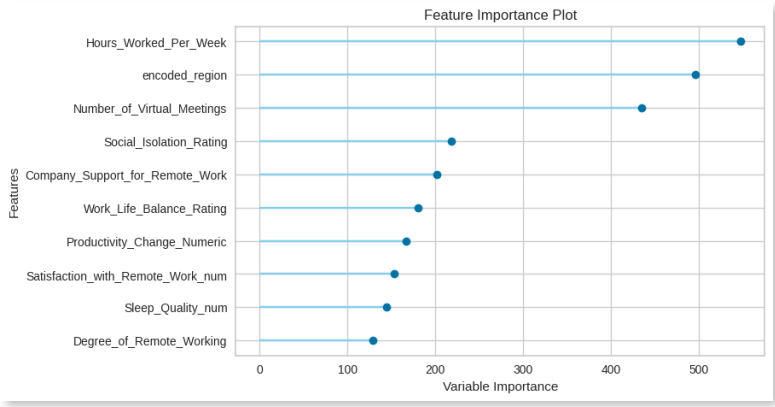




- **SHAP 圖：**地區（encoded_region）是最重要的特徵，再來是每週工時（Hours_Worked_Per_Week）和視訊會議次數（Number_of_Virtual_Meetings）。
- 每週工時與視訊會議次數的紅色點（高值）分布廣，代表每週工時越長、視訊會議次數越多，越可能預測為 1（心理健康問題）。
- **特徵重要性圖：**地區和每週工時是最關鍵的兩個特徵。
- 特徵重要性是依照每個模型內建的特徵重要性計算方法去生成的，而 CatBoost 模型是用 Loss Function-based Importance 作為預設計算方法。

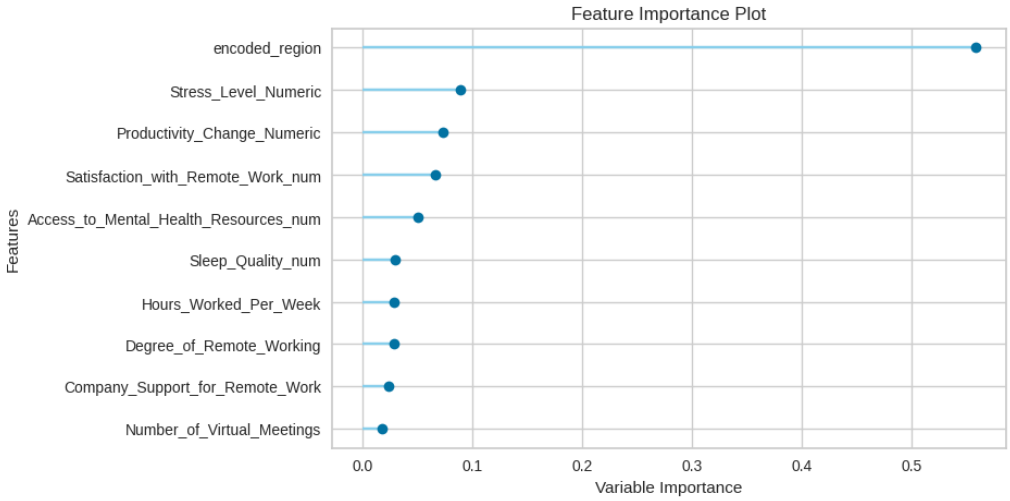
② LightGBM





- **SHAP 圖：**與上一個模型結果類似，地區（encoded_region）影響力最大，其次是每週工時（Hours_Worked_Per_Week）和視訊會議次數（Number_of_Virtual_Meetings）。
- 每周工時與壓力指數對預測結果貢獻顯著，特徵值越高越可能預測為 1。
- **特徵重要性圖：**地區和每週工時的重要性排名前兩位，視訊會議次數和壓力指數次之。
- LightGBM 模型的特徵重要性圖是用 Split Importance 作為預設計算方法。

③ Gradient Boosting Classifier



- **特徵重要性圖：**地區和壓力指數的影響最大，其他特徵貢獻均較小。
- 此模型的特徵重要性圖是用 Mean Decrease in Impurity 作為預設計算方法。

❖ **解釋性比較總結：**

- 三個模型均表現出地區與每週工時為心理健康的關鍵影響因素，但各模型在壓力指數與視訊會議次數的影響強度上有表現出差異。
- LightGBM 在重要特徵識別和解釋方面具有優勢，而 CatBoost 則在捕捉多樣化特徵影響上表現更均衡，Gradient Boosting 適合對正樣本的精準捕捉。

七.結論與未來展望

❖ 結論

1. 遠距工作對心理健康的影響：

- 每週工時、視訊會議次數和地區是心理健康預測中最關鍵的三項特徵；而睡眠品質都在倒數。

2. 模型表現與選擇：

- CatBoost 模型的整體效能（Accuracy、AUC、Recall）最佳，為心理健康狀況預測的首選。
- LightGBM 模型次之，適用於需快速處理大數據集的時候。
- Gradient Boosting Classifier 在捕捉正樣本方面表現優異，但假陽性的比例較高。

3. 改進空間：

- 雖然 CatBoost 模型表現最佳，但看到混淆矩陣後，還是有許多可以改進的部分。

❖ 未來展望

1. 資料擴展：

- 增加更多地區與產業的樣本數據，提升模型對不同文化與工作型態的適用性。
- 探索更多可能影響到心理健康的特徵，例如：個人性格特徵與社交支持程度。

2. 模型優化：

- 調整 Label Encoding 的方式，特別針對像工作型態（Work_Location）這類無順序性的變數，改用 One-Hot Encoding 以避免引入不必要的順序性。

八.參考資料

- 資料來源：https://www.kaggle.com/datasets/iramshahzadi9/remote-work-and-mental-health
- 特徵重要性：
- https://hyades910739.medium.com/%E6%B7%BA%E8%AB%87-tree-model-%E7%9A%84-feature-importance-3de73420e3f2
- https://pycaret.readthedocs.io/en/stable/api/classification.html
- PyCaret：https://github.com/pycaret/pycaret/blob/master/tutorials/Tutorial%20-%20Multiclass%20Classification.ipynb
- ChatGPT
- 圖面來源：https://blog.pinkoi.com/hk/hot-topics/yuanchi_illustrator/