Remote Work & Mental Health

MI32040019 廖廣筑
MI32040022 林士耕

Table of contents

1 資料簡介
來源 & 欄位介紹
數值/類別變數概述

2 資料清洗
缺失值處理
異常值檢查

3 特徵分析
數值變數分佈
類別變數分佈

4 目標變量分析
Label Encoding
Target Encoding

5 特徵 vs 目標變量
特定欄位與心理健康狀況的關係

6 資料不平衡處理
SMOTE 過採樣

7 模型選擇 & 解釋
特定欄位與心理健康狀況的關係

8 結論與未來展望
結合期中 & 期末的結論

1

資料簡介

About the Dataset

- Data Source：Kaggle
- Number of Observations：5000
- Total Columns：20

- Numerical Columns：7
- Categorical Columns：13

- 目標變數：Mental_Health_Condition

Numerical Columns

Column	Explanation	(Min, Median, Max)
Age	年齡	(22 , 41 , 60)
Years_of_Experience	工作經驗（年）	(1 , 18 , 35)
Hours_Worked_Per_Week	每週工時（時）	(20 , 40 , 60)
Number_of_Virtual_Meetings	視訊會議次數	(0 , 8 , 15)
Work_Life_Balance_Rating	工作與生活平衡評級	(1 , 3 , 5)
Social_Isolation_Rating	社會孤立評分(感到孤立程度)	(1 , 3 , 5)
Company_Support_for_Remote_Work	公司對遠距工作的支持	(1 , 3 , 5)

Categorical Columns (1)

Column	Explanation	Data Types
Employee_ID	員工的唯一識別碼	EMP0001～EMP5000
Gender	性別	Non-binary / Female / Male / Prefer not to say
Job_Role	目前的工作崗位	HR / Data Scientist / Software Engineer / Sales / Marketing / Designer / Project Manager
Industry	產業	Healthcare / IT / Education / Finance / Consulting / Manufacturing / Retail
Work_Location	工作型態	Hybrid(混合工作) / Remote(遠端工作) / Onsite(現場工作)
Stress_Level	壓力指數(壓力程度)	Medium / High / Low

Categorical Columns (2)

Column	Explanation	Data Types
Access_to_Mental_Health_Resources	是否獲取心理健康資源	Yes / No
Productivity_Change	生產力是否改變	Decrease(下降) / Increase(上升) / No Chan(不變)
Satisfaction_with_Remote_Work	對遠端工作的滿意度	Satisfied(滿意) / Unsatisfied(不滿意) / Neutral(中性)
Physical_Activity	身體活動	Weekly / Daily / nan
Sleep_Quality	睡眠品質	Good / Poor / Average
Region	地區	Europe / Asia / North America / South America / Oceania / Africa
Mental_Health_Condition	心理健康狀況	Depression(憂鬱) / Anxiety(焦慮) / Burnout(職業倦怠)

2

資料清洗

Check for Missing Values

有缺失的變數	個數（比例）	處理方式
Mental_Health_Condition	1196（23.92%）	Nan 填入 None 表示沒有其他三種心理狀態
Physical_Activity	1629（32.58%）	Nan 值全部填入 0 當成該員工沒有運動習慣

3、特徵分析

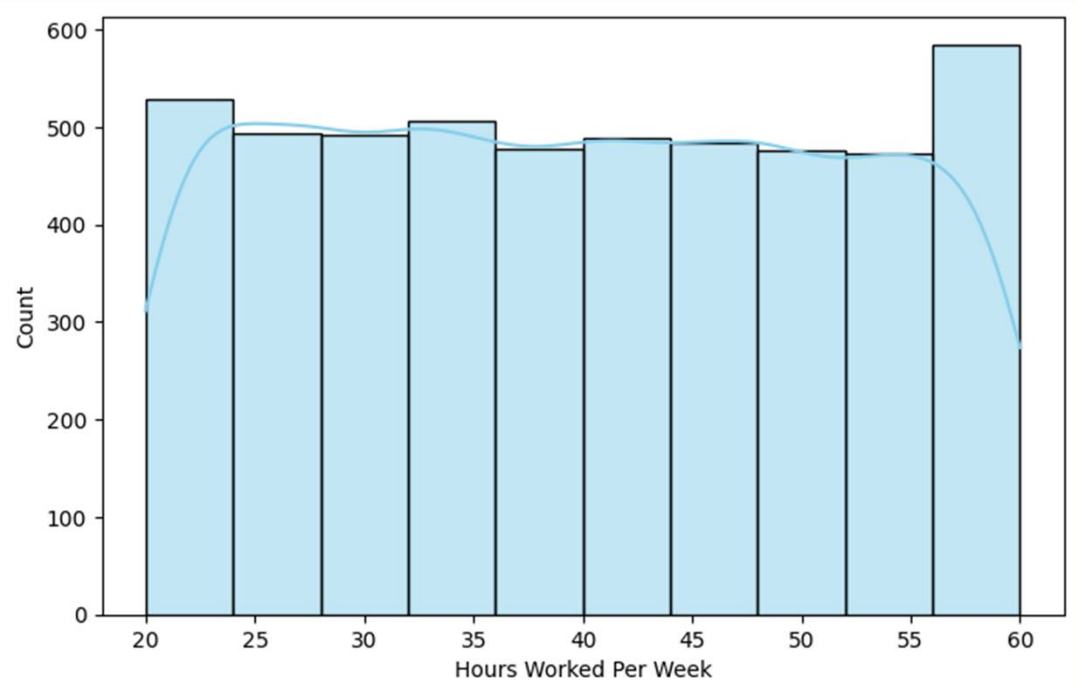
Numerical

- Hours Worked Per Week
- Number of Virtual Meetings
- Work Life Balance Rating
- Social Isolation Rating
- Social Isolation Rating

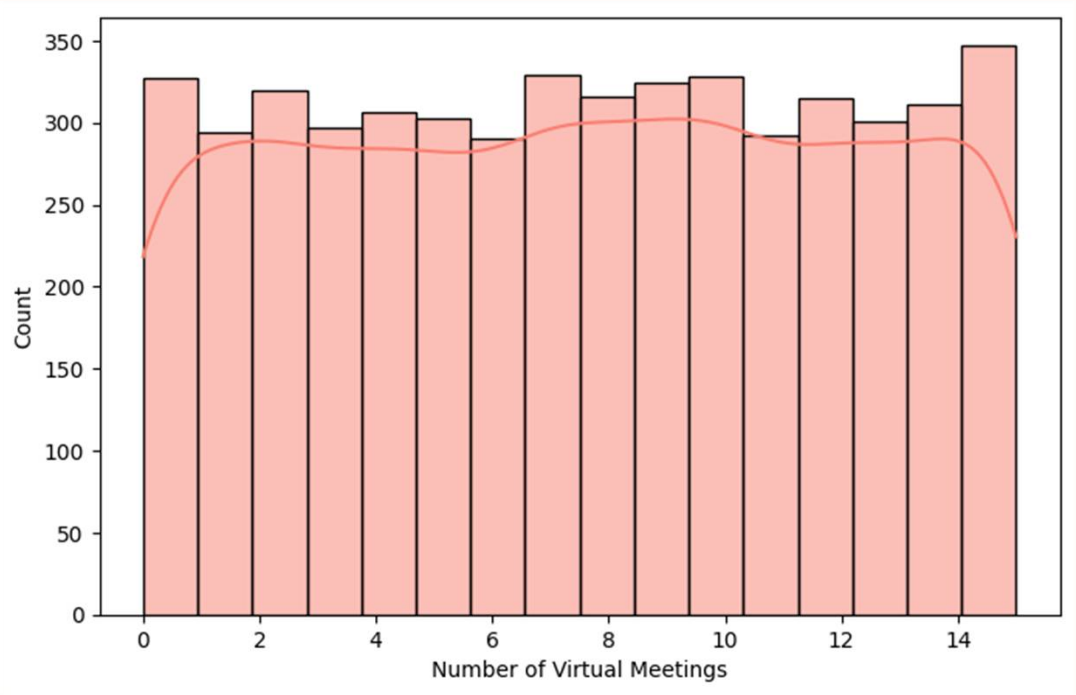
Freepik

Distribution of Numerical Features

- Hours Worked Per Week

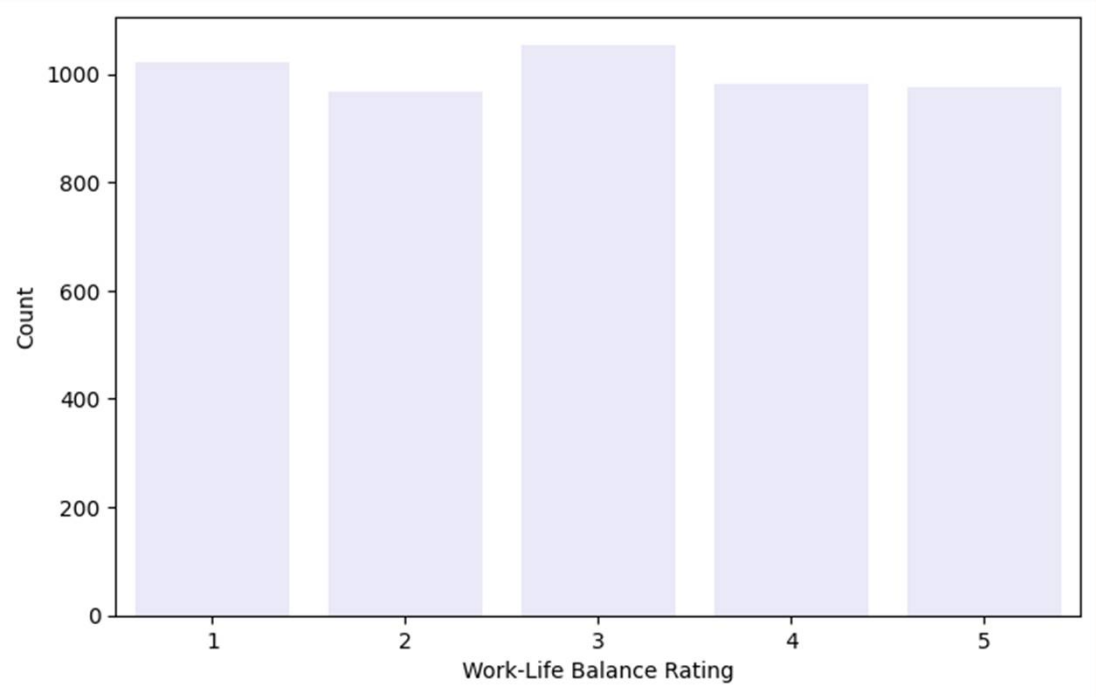


- Number of Virtual Meetings

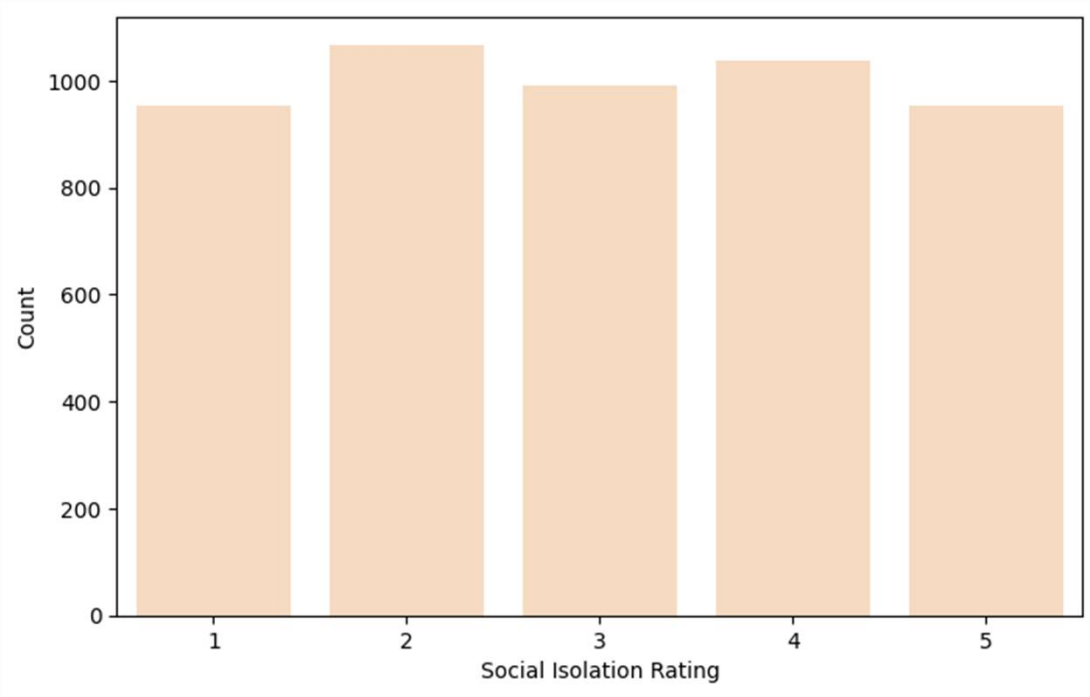


Distribution of Numerical Features

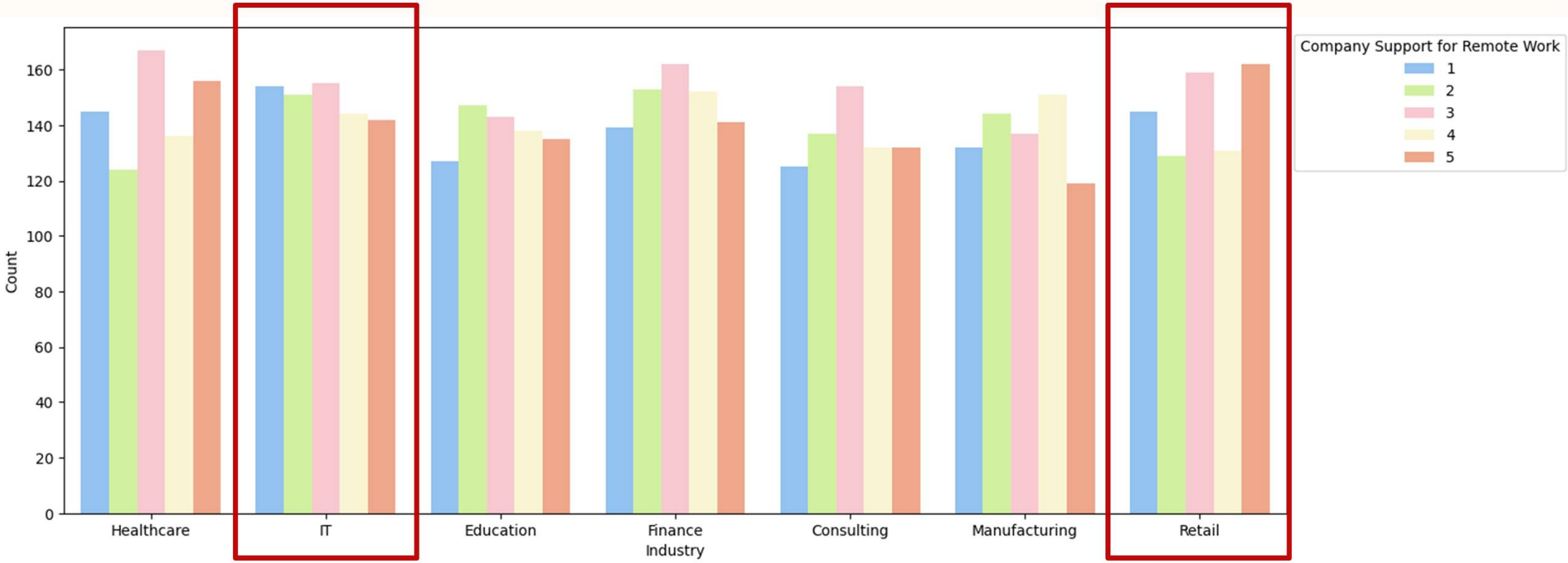
- Work-Life Balance Rating



- Social Isolation Rating



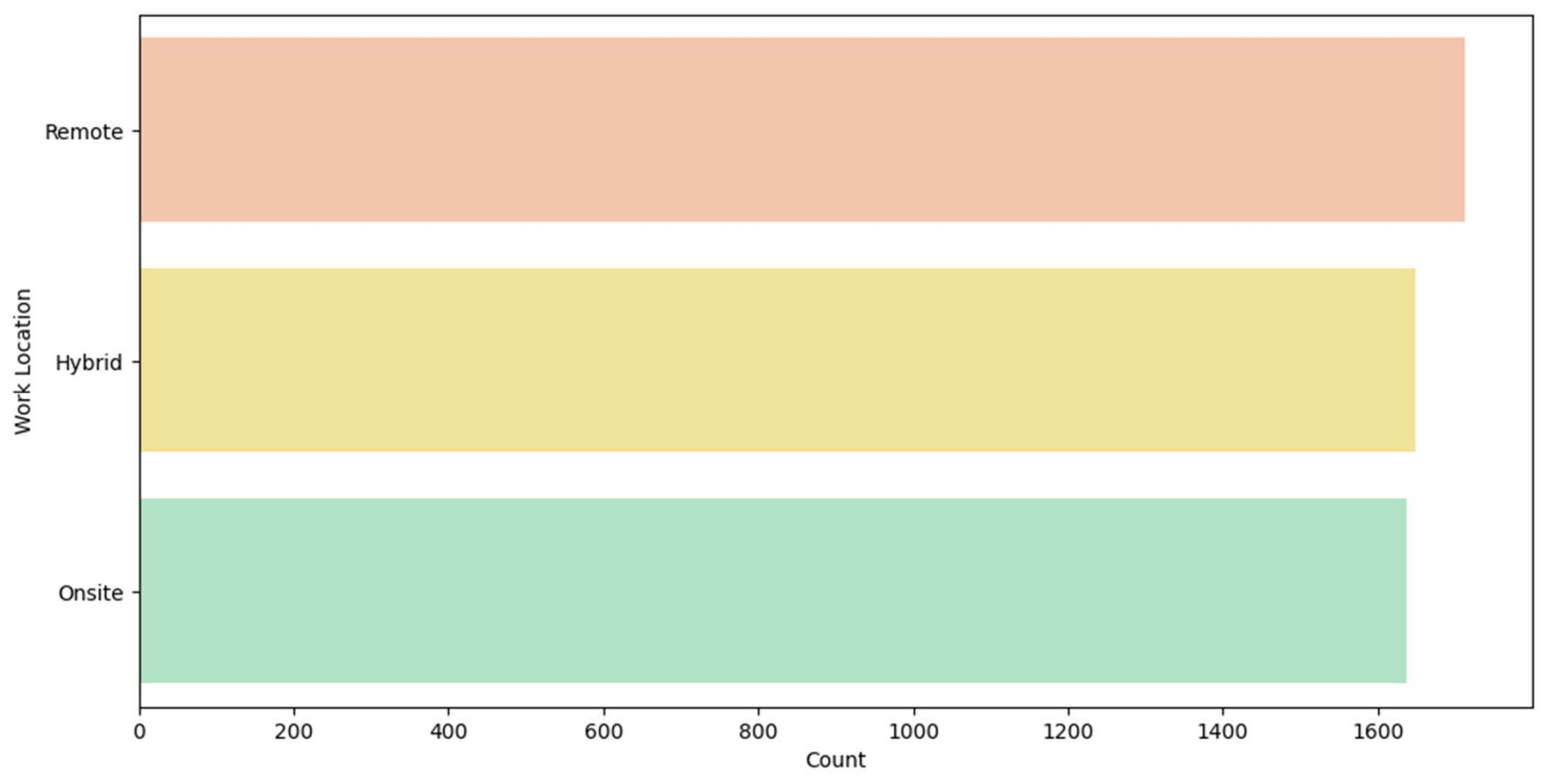
Industry vs. Company Support for Remote Work



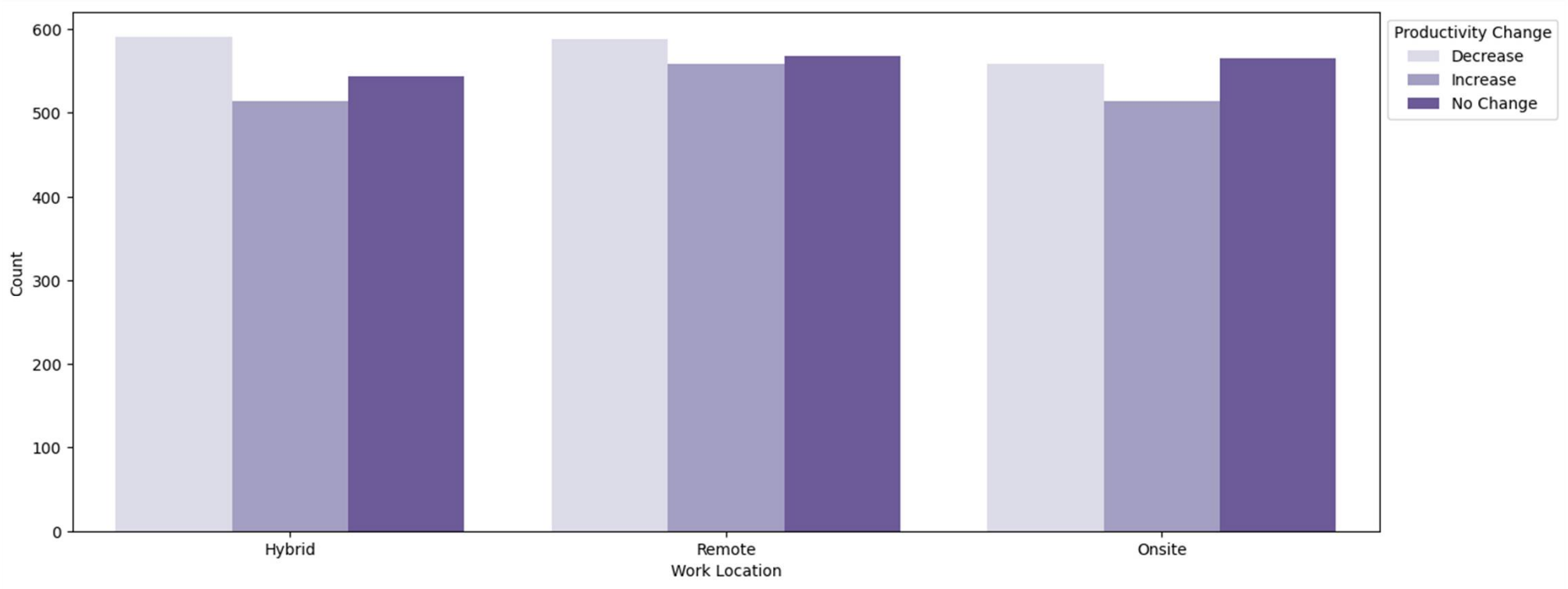
Categorical

- Job Role
- Work Location
- Stress Level
- Mental Health Condition
- Sleep Quality
- Access to Mental Health Resources
- Productivity Change
- Physical Activity

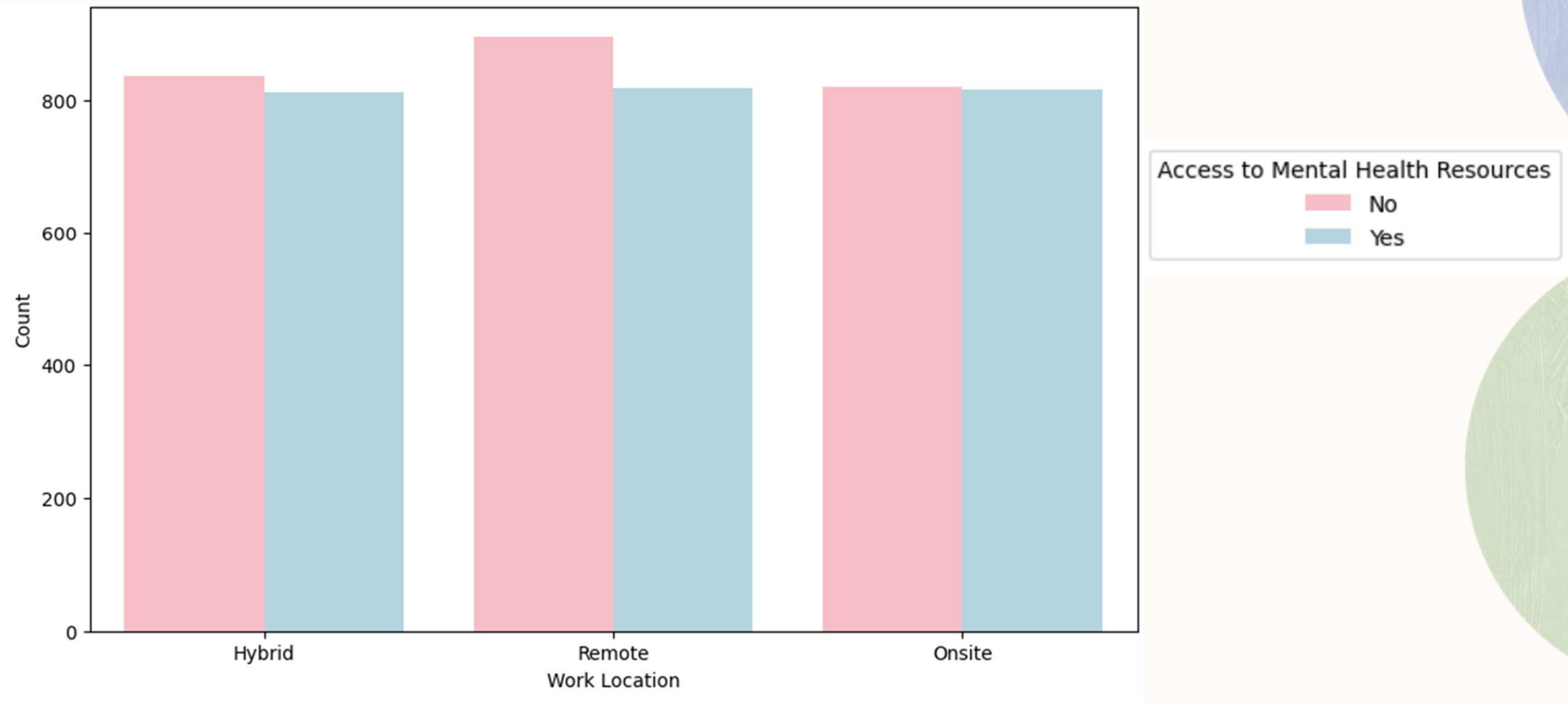
Work Location Distribution



Work Location vs. Productivity Change

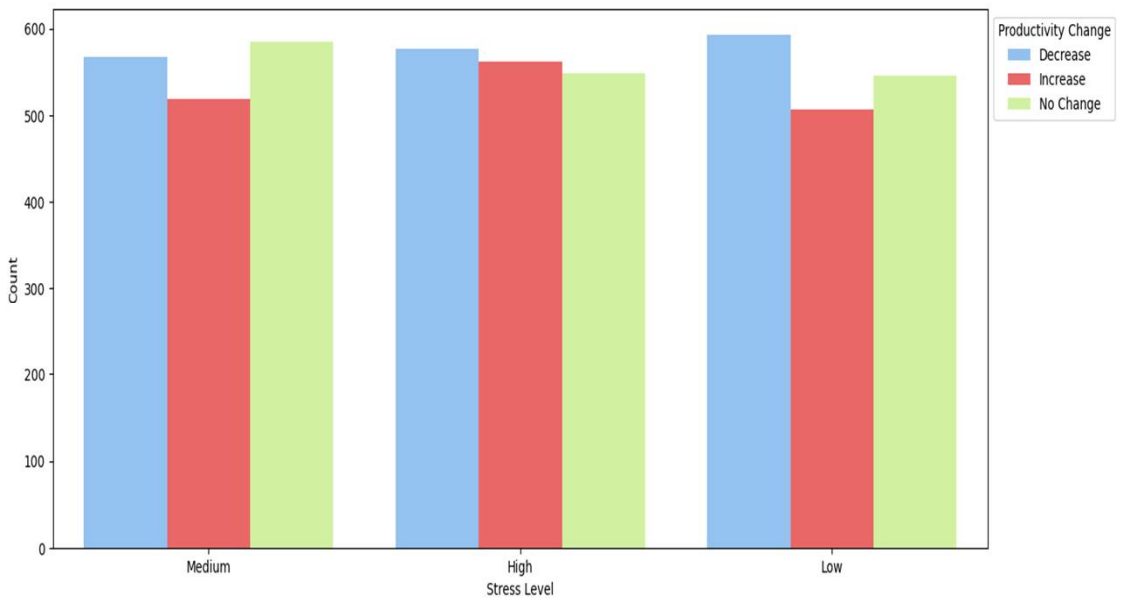


Work Location vs. Access to Mental Health Resources

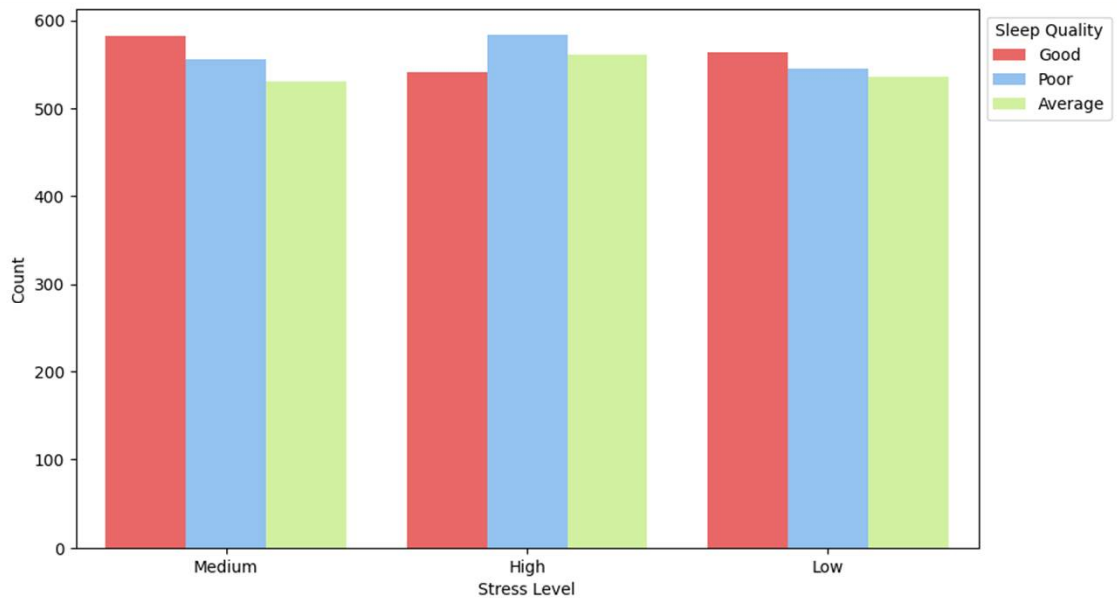


Stress Level

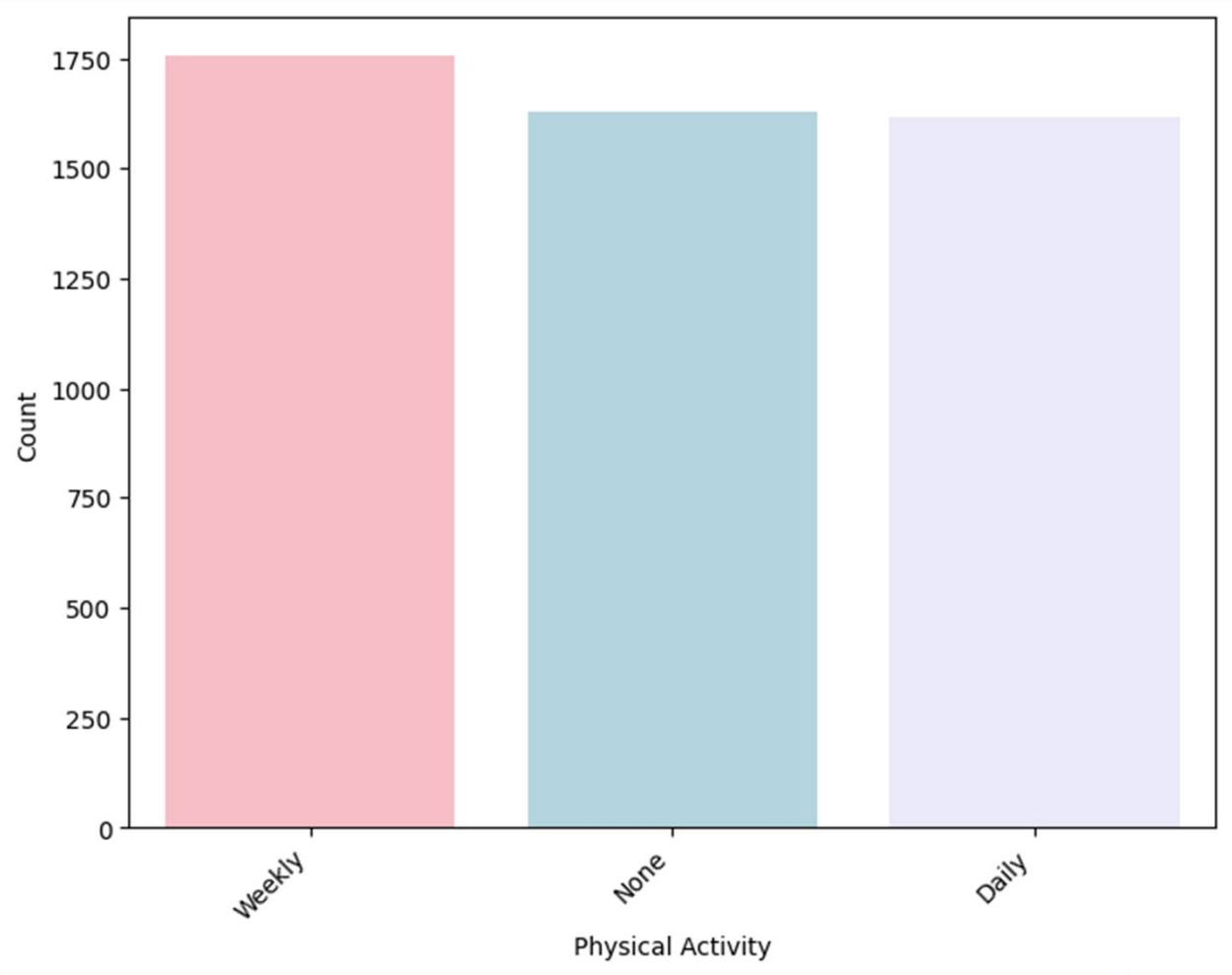
- Productivity_Change



- Sleep Quality

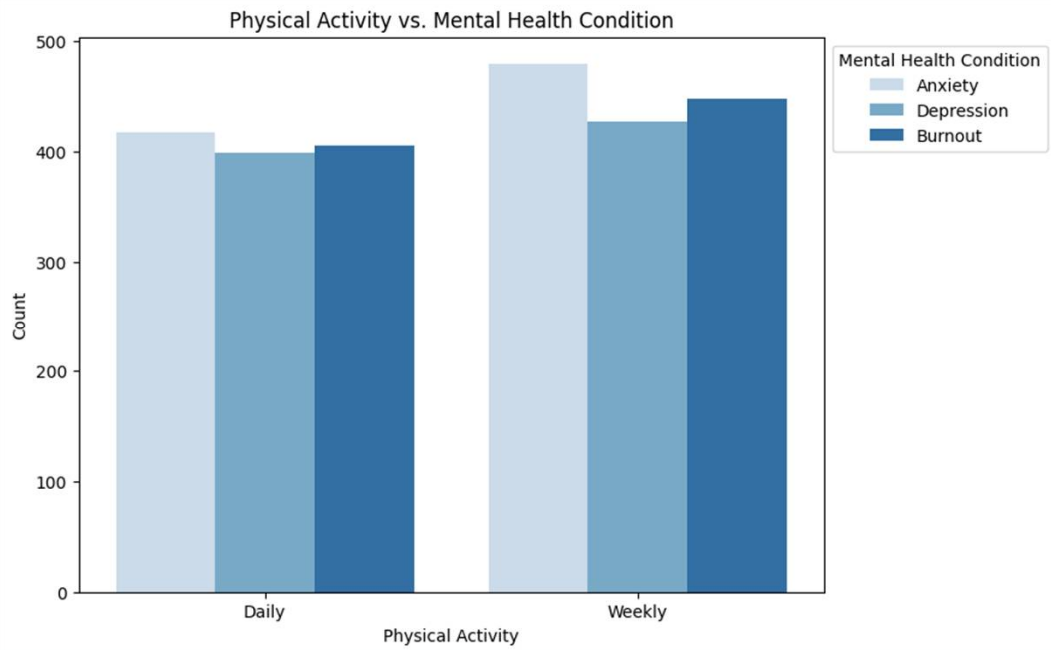


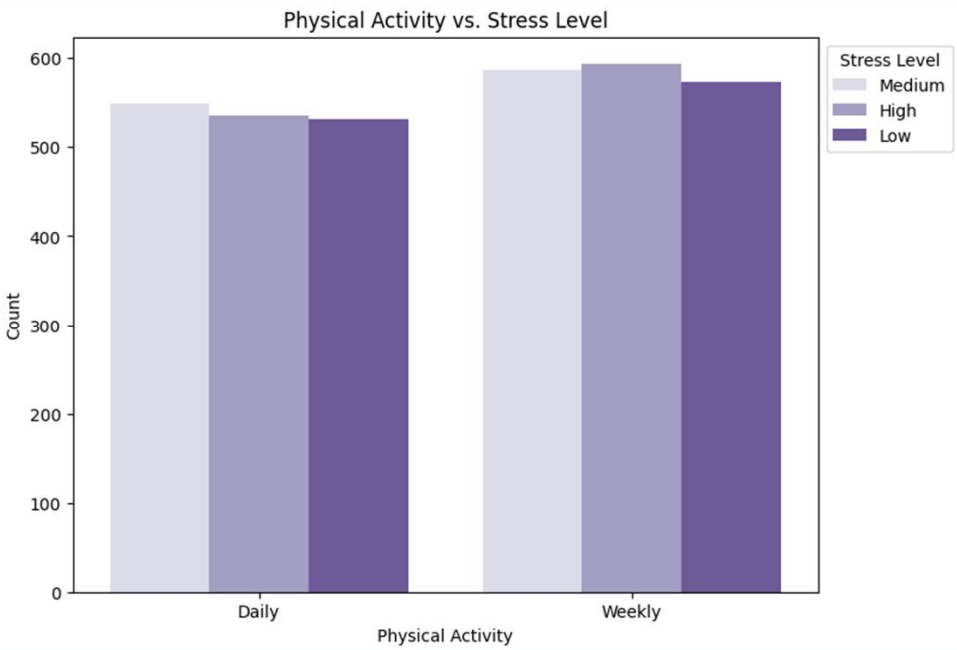
Physical Activity



Physical Activity

- Mental Health Condition
- Stress Level

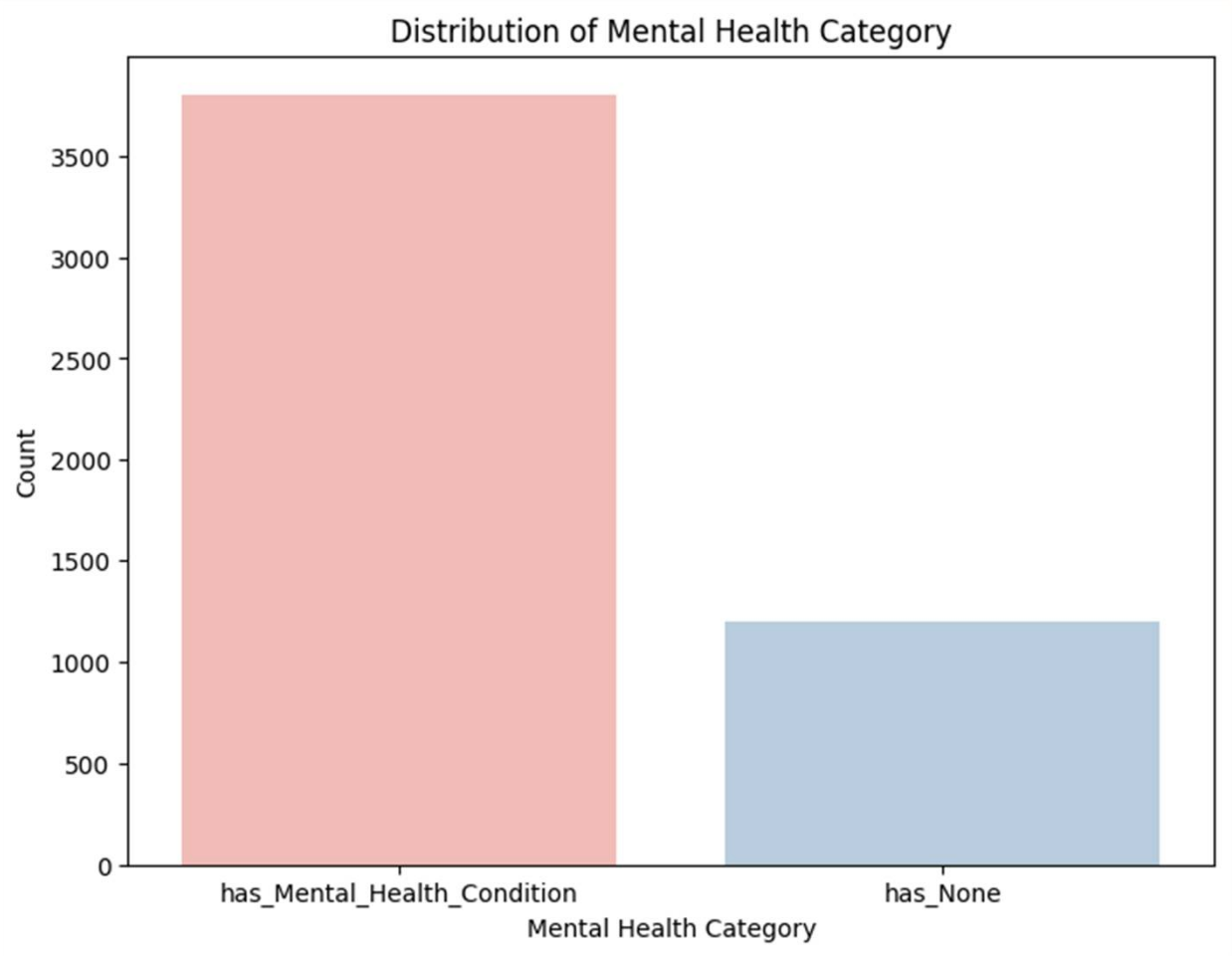




4

目標變量分析

目標變數：心理健康狀況


Distribution of Mental Health Category
Count
3500
3000
2500
2000
1500
1000
500
0
has_Mental_Health_Condition
has_None
Mental Health Category

Label Encoding

Before		After	
Column	Data Types	Column	Data Types
心理健康狀況 Mental_Health_Condition	has_Mental_Health_Condition / has_None	Mental_Health_Category_num	1 / 0
壓力指數 Stress_Level	Low / Medium / High	Stress_Level_Numeric	-1 / 0 / 1
是否獲取心理健康資源 Access_to_Mental_Health_Resources	Yes / No	Access_to_Mental_Health_Resources_Numeric	1 / 0
工作型態 Work_Location	Onsite / Hybrid / Remote	Degree_of_Remote_Working	0 / 1 / 2

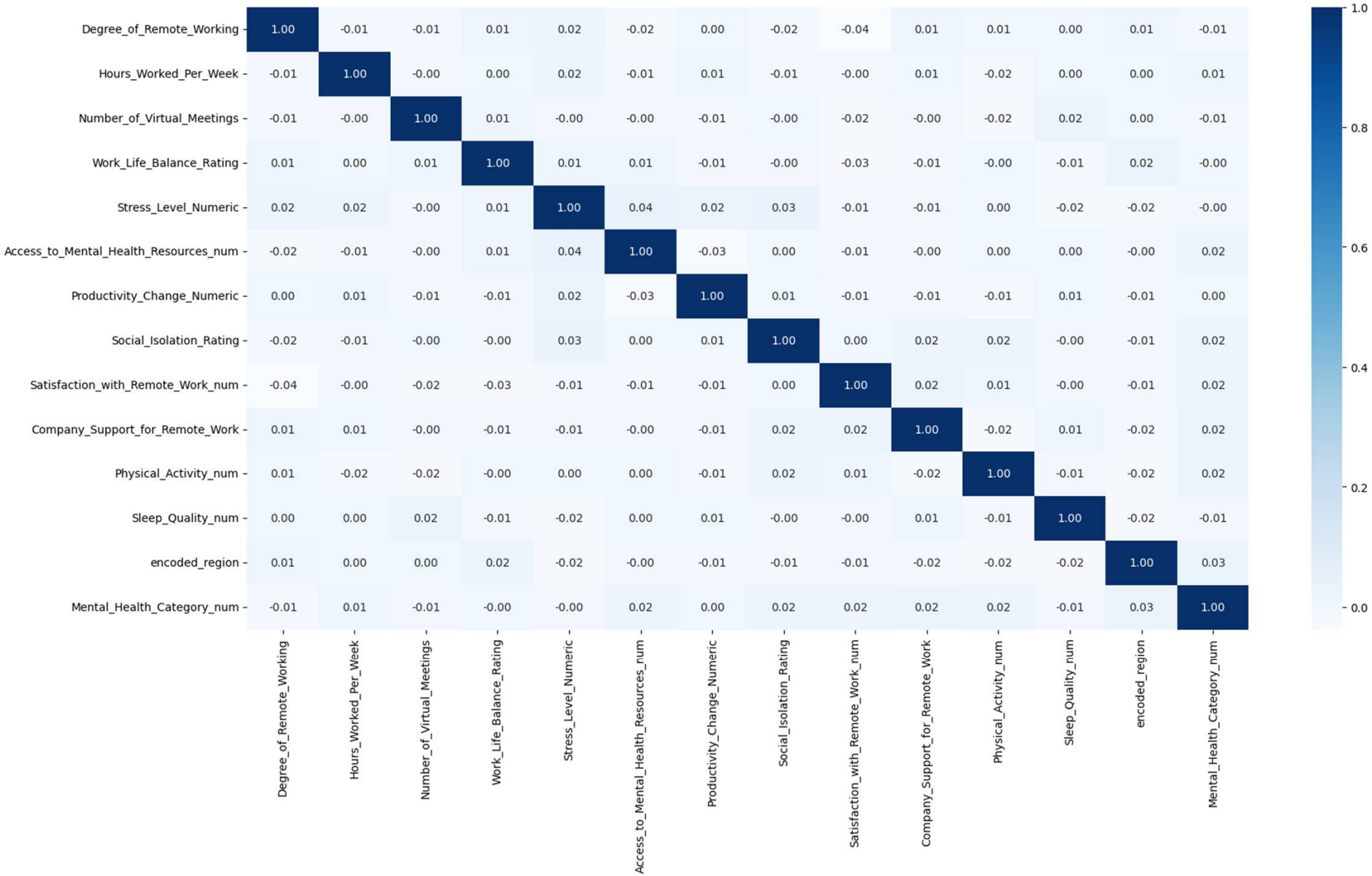
Label Encoding

Before		After	
Column	Data Types	Column	Data Types
生產力 Productivity_Change	Decrease / No Change / Increase	Productivity_Change_Numeric	-1 / 0 / 1
對遠距工作的滿意度 Satisfaction_with_Remote_Work	Unsatisfied / Neutral / Satisfied	Satisfaction_with_Remote_Work_num	-1 / 0 / 1
身體活動 Physical_Activity	None / Daily / Weekly	Physical_Activity_num	0 / 1 / 2
睡眠品質 Sleep_Quality	Poor / Average / Good	Sleep_Quality_num	-1 / 0 / 1

Target Encoding

Before		After	
Column	Data Types	Column	Data Types
Region	Europe / Asia / North America / South America / Oceania / Africa	Encoded_region	0.7511627906976744, 0.7520184544405998, 0.7567567567567568, 0.7593712212817413, 0.7607142857142857, 0.7852834740651388.

Correlation Heatmap

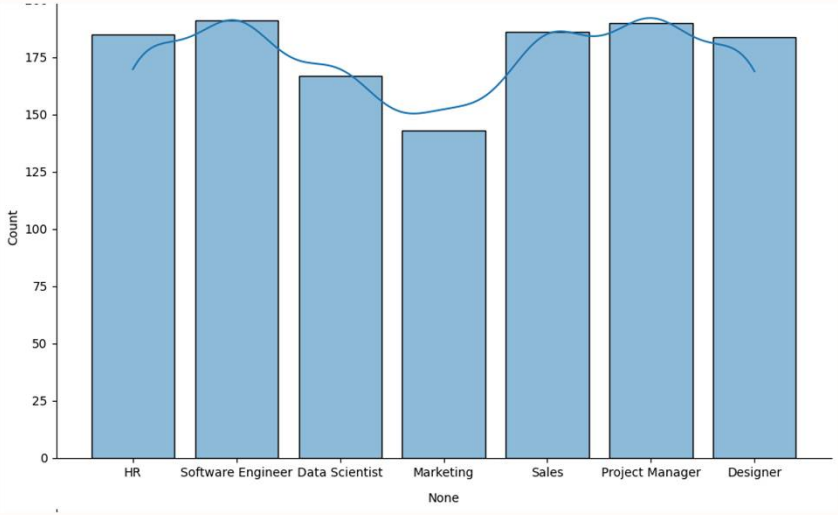

Degree_of_Remote_Working 1.00 -0.01 -0.01 0.01 0.02 -0.02 0.00 -0.02 -0.04 0.01 0.01 0.00 0.01 -0.01
Hours_Worked_Per_Week -0.01 1.00 -0.00 0.00 0.02 -0.01 0.01 -0.01 -0.00 0.01 -0.02 0.00 0.00 0.01
Number_of_Virtual_Meetings -0.01 -0.00 1.00 0.01 -0.00 -0.00 -0.01 -0.00 -0.02 -0.00 -0.02 0.02 0.00 -0.01
Work_Life_Balance_Rating 0.01 0.00 0.01 1.00 0.01 0.01 -0.01 -0.00 -0.03 -0.01 -0.00 -0.01 0.02 -0.00
Stress_Level_Numeric 0.02 0.02 -0.00 0.01 1.00 0.04 0.02 0.03 -0.01 -0.01 0.00 -0.02 -0.02 -0.00
Access_to_Mental_Health_Resources_num -0.02 -0.01 -0.00 0.01 0.04 1.00 -0.03 0.00 -0.01 -0.00 0.00 0.00 -0.00 0.02
Productivity_Change_Numeric 0.00 0.01 -0.01 -0.01 0.02 -0.03 1.00 0.01 -0.01 -0.01 -0.01 0.01 -0.01 0.00
Social_Isolation_Rating -0.02 -0.01 -0.00 -0.00 0.03 0.00 0.01 1.00 0.00 0.02 0.02 -0.00 -0.01 0.02
Satisfaction_with_Remote_Work_num -0.04 -0.00 -0.02 -0.03 -0.01 -0.01 -0.01 0.00 1.00 0.02 0.01 -0.00 -0.01 0.02
Company_Support_for_Remote_Work 0.01 0.01 -0.00 -0.01 -0.01 -0.00 -0.01 0.02 0.02 1.00 -0.02 0.01 -0.02 0.02
Physical_Activity_num 0.01 -0.02 -0.02 -0.00 0.00 0.00 -0.01 0.02 0.01 -0.02 1.00 -0.01 -0.02 0.02
Sleep_Quality_num 0.00 0.00 0.02 -0.01 -0.02 0.00 0.01 -0.00 -0.00 0.01 -0.01 1.00 -0.02 -0.01
encoded_region 0.01 0.00 0.00 0.02 -0.02 -0.00 -0.01 -0.01 -0.01 -0.02 -0.02 -0.02 1.00 0.03
Mental_Health_Category_num -0.01 0.01 -0.01 -0.00 -0.00 0.02 0.00 0.02 0.02 0.02 0.02 -0.01 0.03 1.00
Degree_of_Remote_Working
Hours_Worked_Per_Week
Number_of_Virtual_Meetings
Work_Life_Balance_Rating
Stress_Level_Numeric
Access_to_Mental_Health_Resources_num
Productivity_Change_Numeric
Social_Isolation_Rating
Satisfaction_with_Remote_Work_num
Company_Support_for_Remote_Work
Physical_Activity_num
Sleep_Quality_num
encoded_region
Mental_Health_Category_num
1.0
0.8
0.6
0.4
0.2
0.0

5

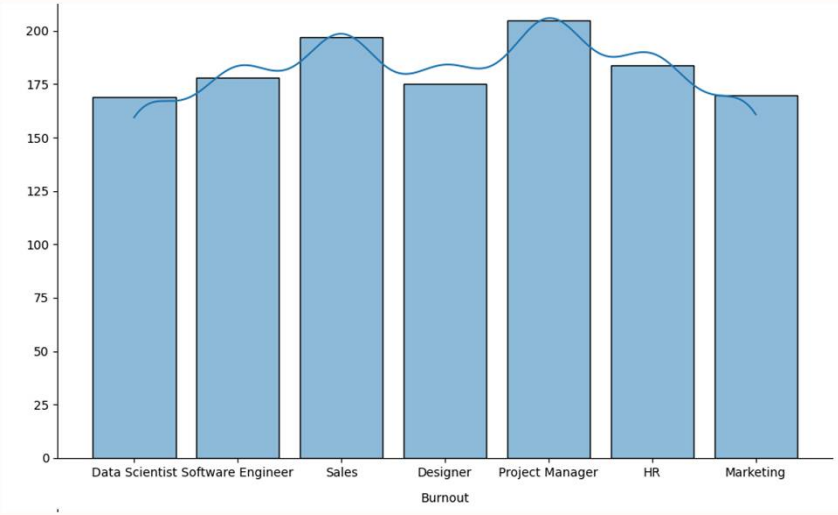
特徵 vs 目標變量

Mental Health Condition vs. Job Role

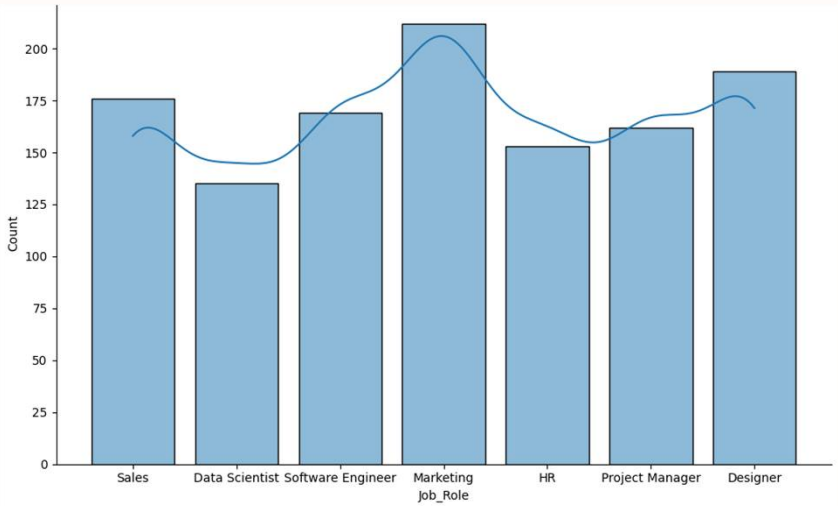
Depression



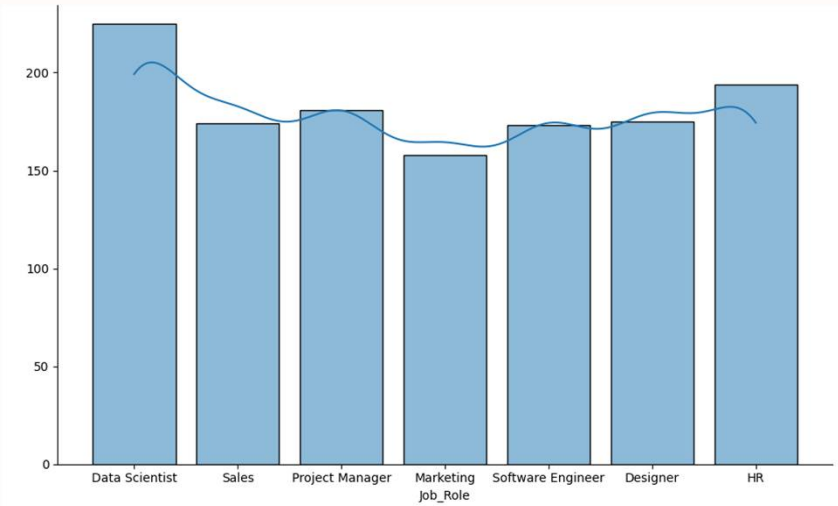
Anxiety



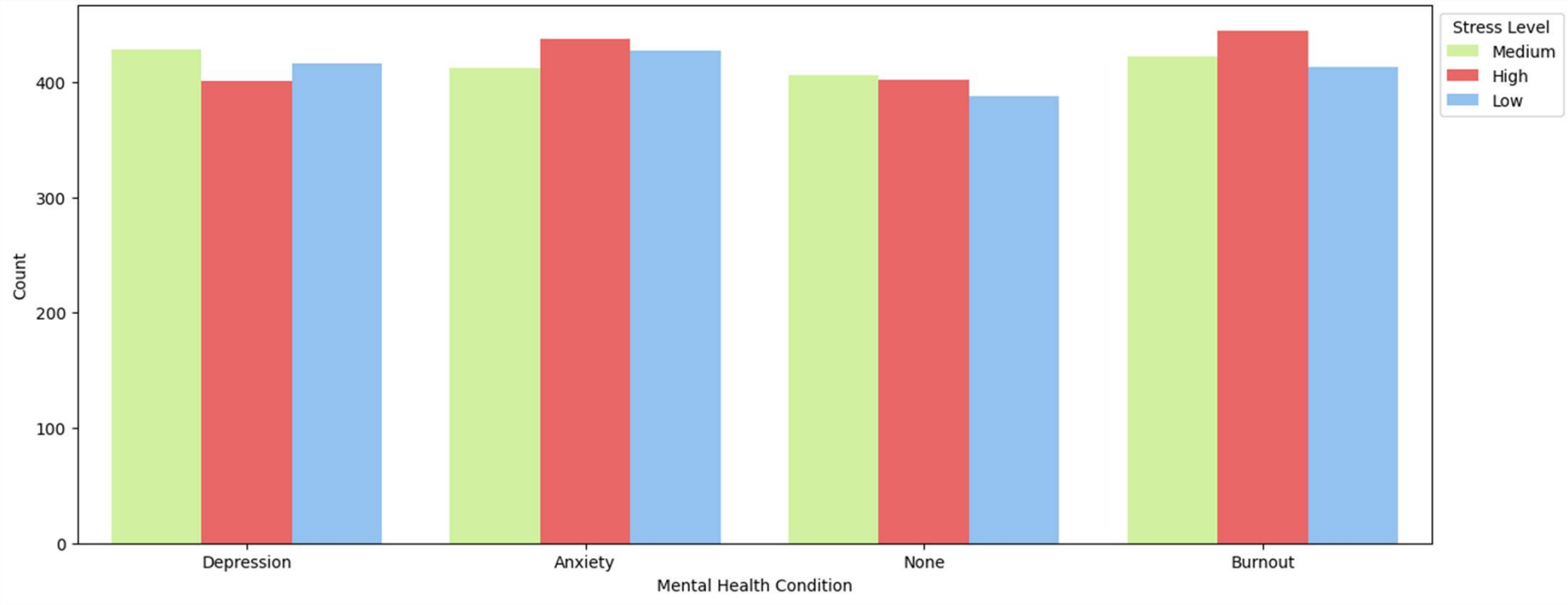
None



Burnout



Mental Health Condition vs. Stress Level



EDA Report 結論

1. 身體活動頻率與壓力指數
2. 產業類型與公司對遠距工作的支持之差異
3. 睡眠品質與壓力水準的潛在相關性

- 各個變數之間的線性關係不強
- 數據不平衡：在某些欄位（Mental_Health_Condition / Work_Location）中，某些類別的樣本數較少，這可能導致分類器在模型訓練時偏向樣本數多的類別。

改善方向

- 數據中潛在資料不平衡的問題，需要在後續建立模型中加以處理，以確保模型的準確性。

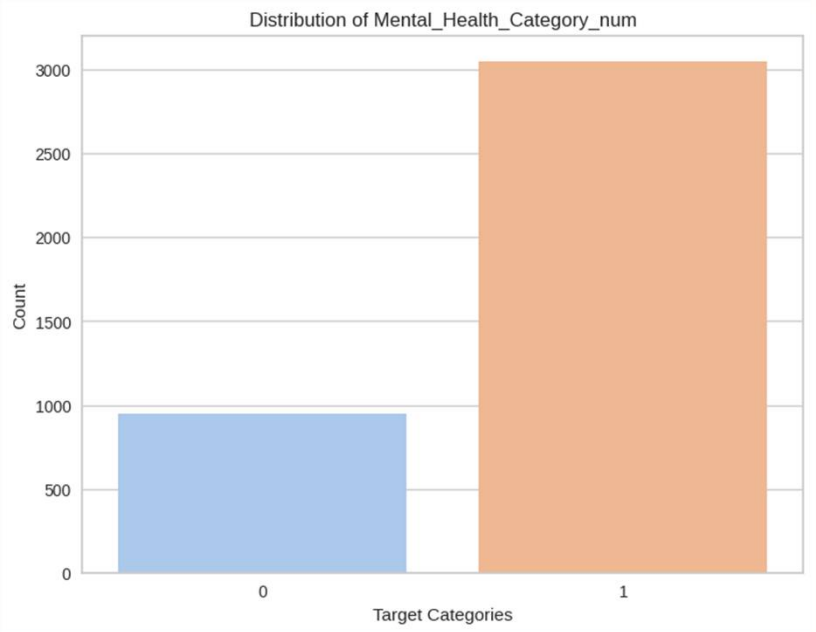
- 在接下來的資料分析中，我們會對欄位進行特徵工程，並挑選非線性模型。在模型建構中考慮這些變量對心理健康的潛在影響，以便為心理健康管理提供更精確的決策。

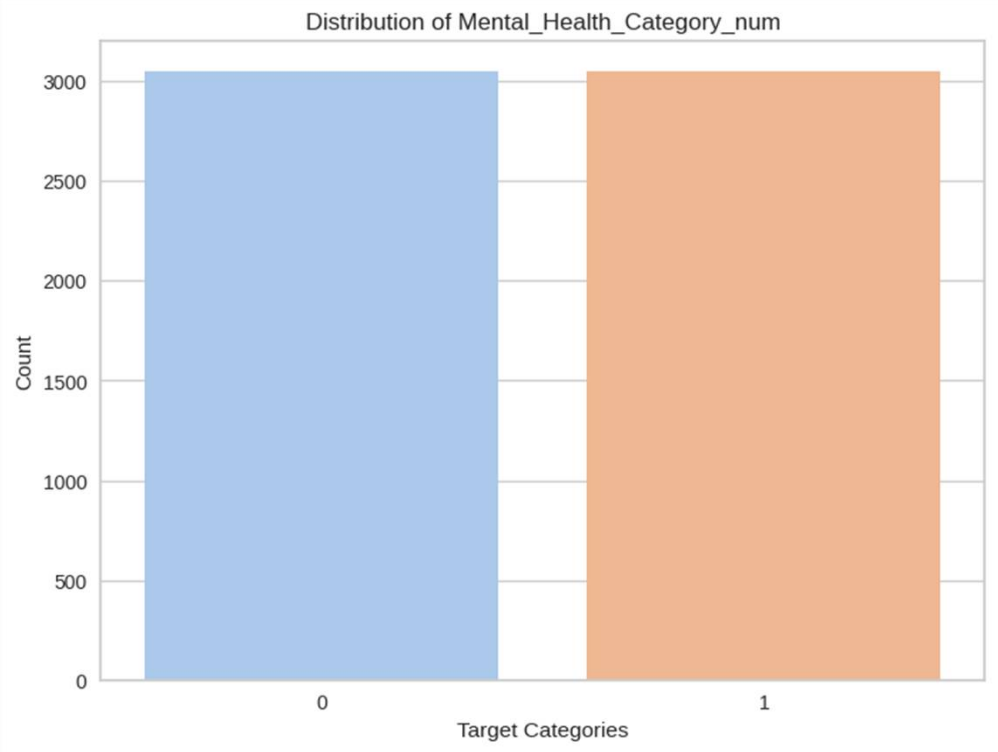
6

資料不平衡處理

SMOTE 過採樣

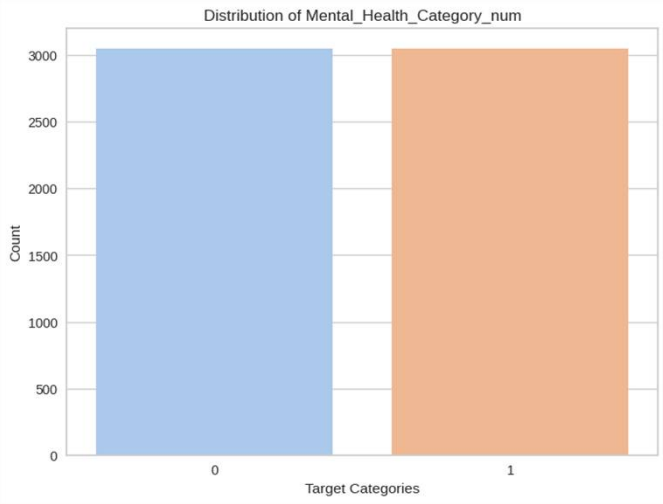
因為在目標變數中，0的數量非常少，所以我們考慮用SMOTE的方法去平衡訓練集數據。





Compare Model

模型	Accuracy	AUC	F1
catboost	0.8136	0.8604	0.8308
lightgbm	0.8089	0.8569	0.8258
gbc	0.8061	0.8461	0.8294
xgboost	0.8036	0.8624	0.8183



7

模型選擇 & 解釋

模型選擇

CatBoost Classifier

LightGradient BoostingMachine

Gradient Boosting Classifier

Extreme Gradient Boosting

Random Forest Classifier

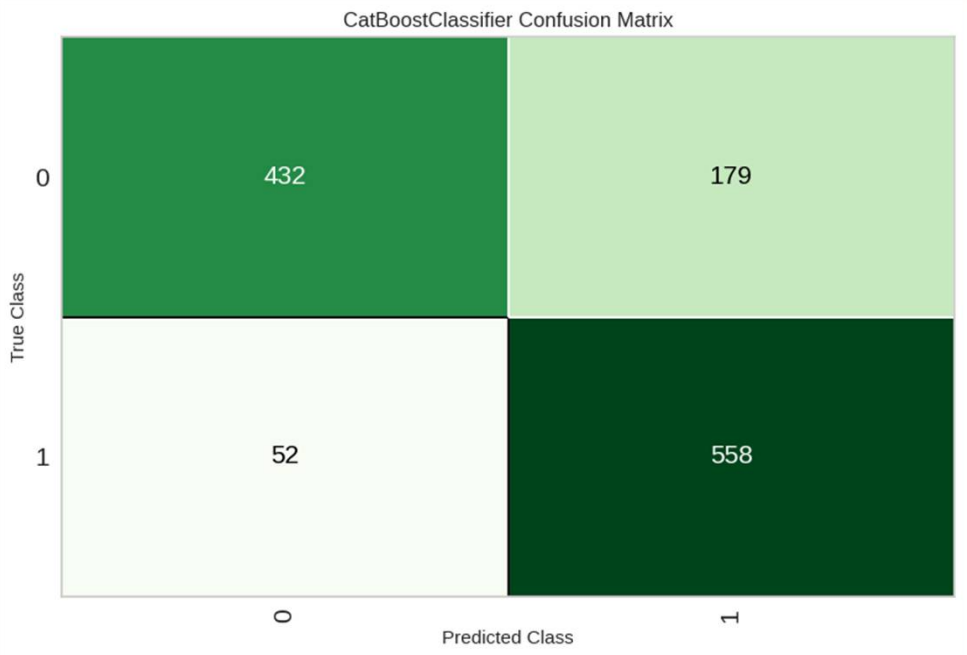
Extra Tree Classifier

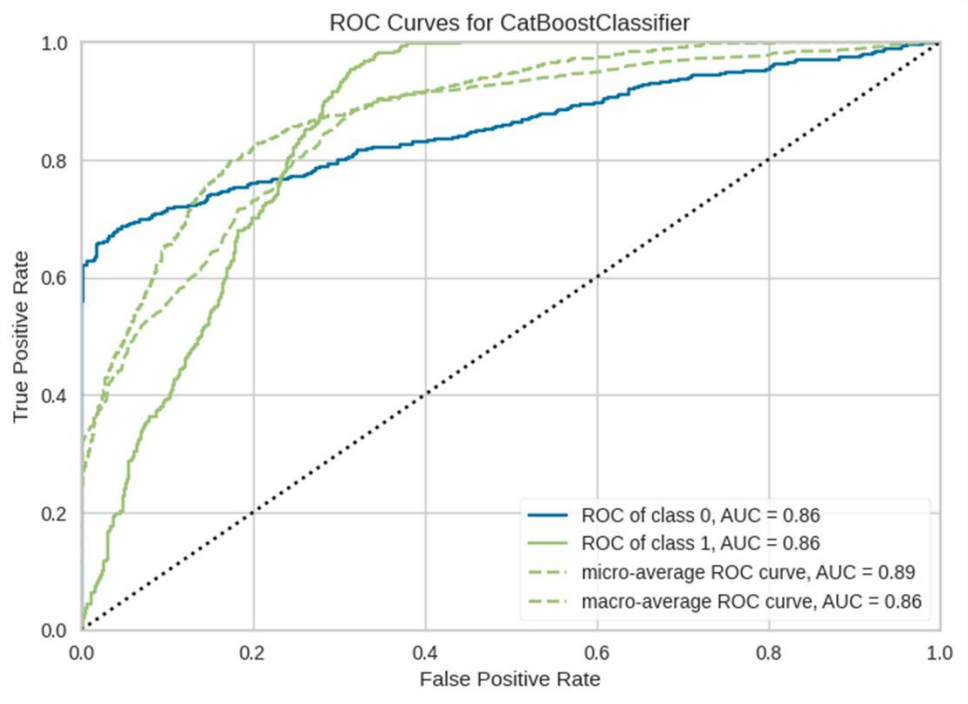
CatBoost Classifier

- CatBoost 是梯度提升樹模型框架
- 特點：對類別特徵的直接支援，也允許字串類型的特徵進行模型訓練
- 以下是此模型在測試集上的表現：

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
CatBoost Classifier	0.7120	0.4810	0.9255	0.7500	0.8286	-0.0131	-0.0166

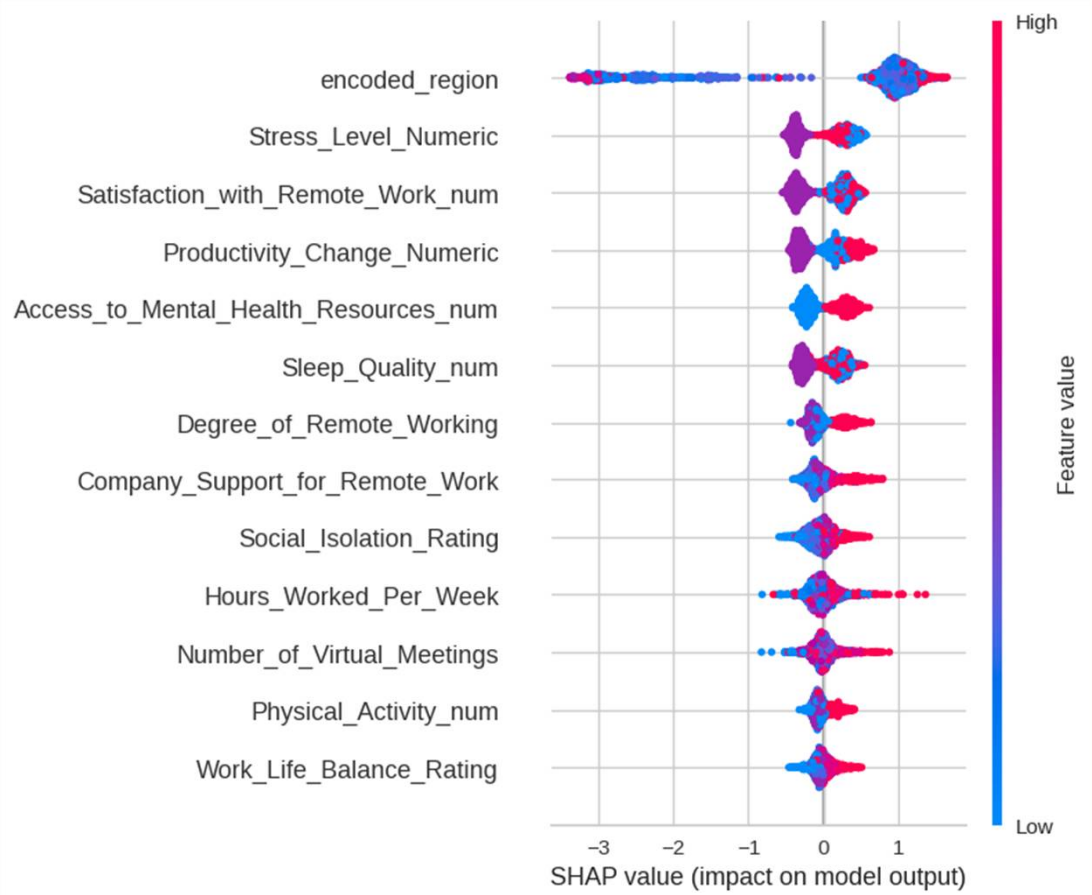
混淆矩陣 & ROC曲線



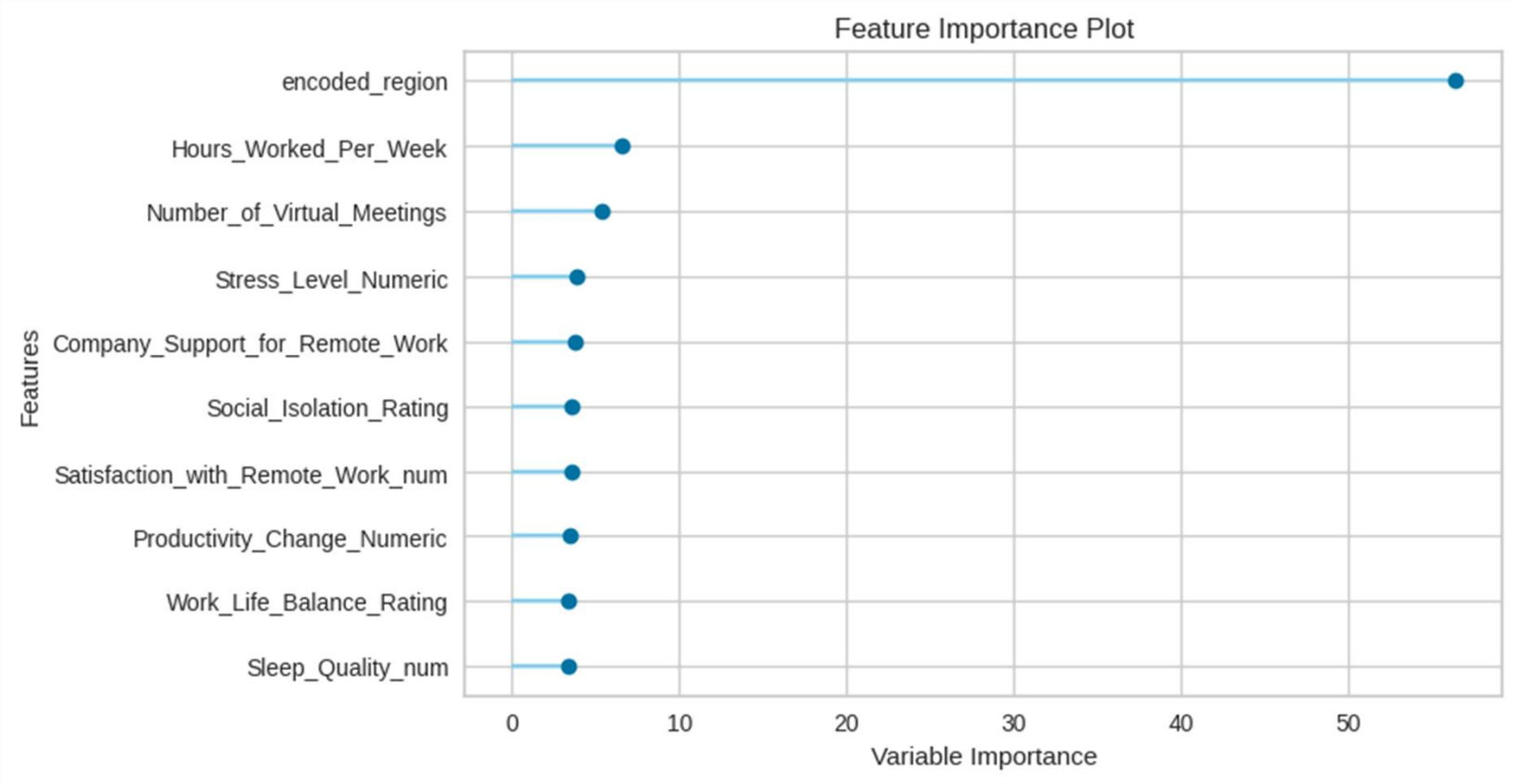


*在測試集上進行

SHAP


encoded_region
Stress_Level_Numeric
Satisfaction_with_Remote_Work_num
Productivity_Change_Numeric
Access_to_Mental_Health_Resources_num
Sleep_Quality_num
Degree_of_Remote_Working
Company_Support_for_Remote_Work
Social_Isolation_Rating
Hours_Worked_Per_Week
Number_of_Virtual_Meetings
Physical_Activity_num
Work_Life_Balance_Rating
−3
−2
−1
0
1
SHAP value (impact on model output)
High
Feature value
Low

特徵重要性圖

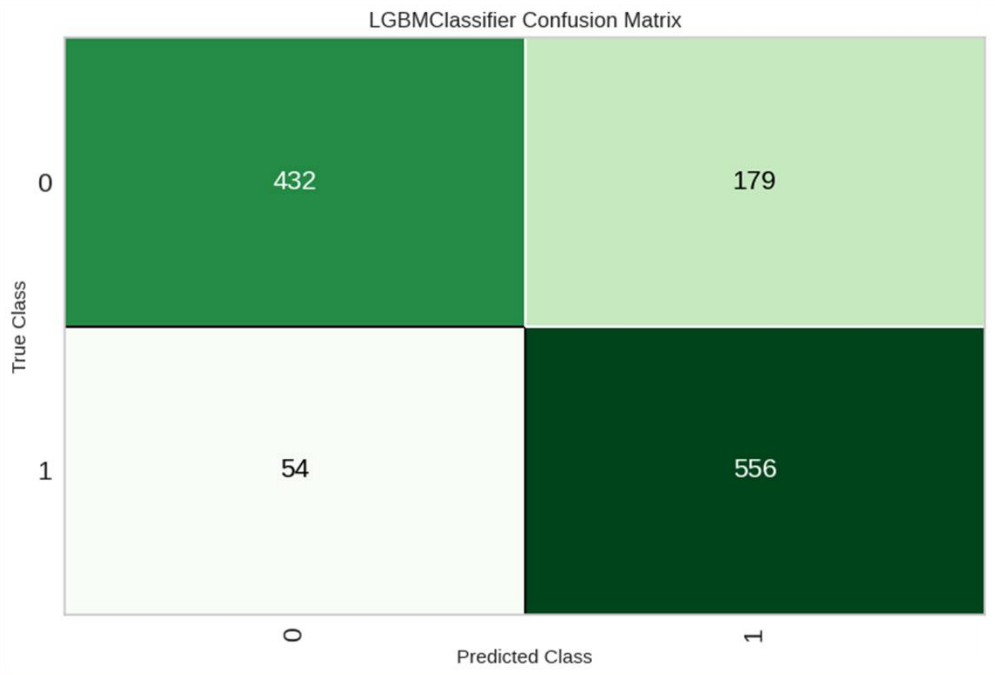

Feature Importance Plot
encoded_region
Hours_Worked_Per_Week
Number_of_Virtual_Meetings
Stress_Level_Numeric
Company_Support_for_Remote_Work
Social_Isolation_Rating
Satisfaction_with_Remote_Work_num
Productivity_Change_Numeric
Work_Life_Balance_Rating
Sleep_Quality_num
Features
0
10
20
30
40
50
Variable Importance

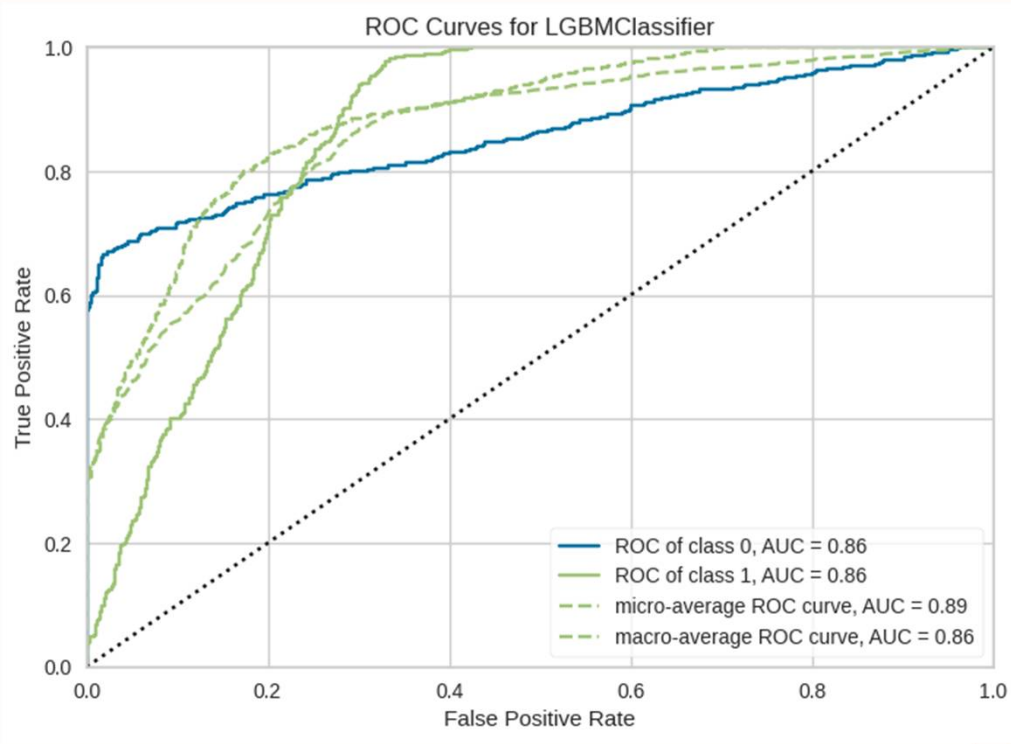
Light Gradient Boosting Machine

- 是一個基於決策樹演算法的分布式梯度提升架構
- 提供快速、高效、低記憶體、高準確度、支援並行和大規模資料處理的工具
- 特點：對類別特徵的直接支援，也允許字串類型的特徵進行模型訓練
- 以下是此模型在測試集上的表現：

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Light Gradient Boosting Machine	0.7270	0.4906	0.9388	0.7567	0.8380	0.0311	0.0406

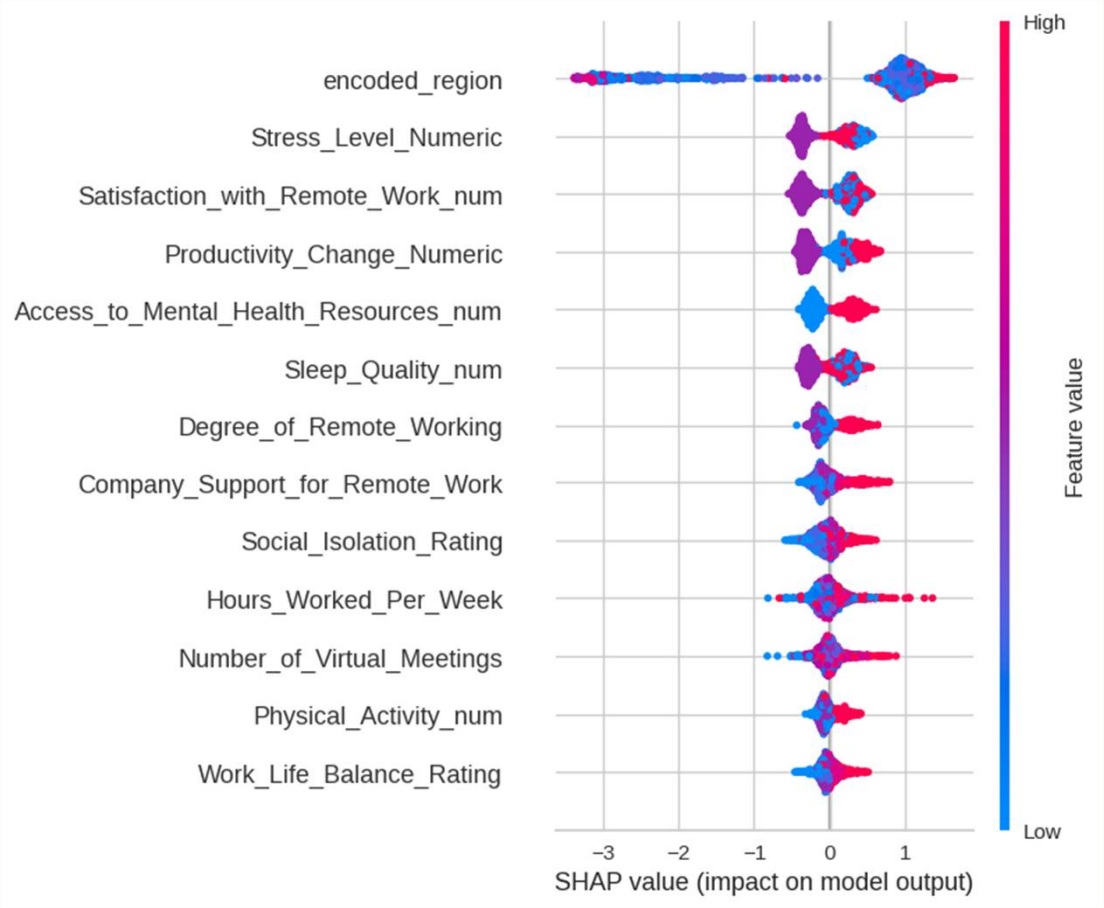
混淆矩陣 & ROC曲線





*在測試集上進行

SHAP


encoded_region
Stress_Level_Numeric
Satisfaction_with_Remote_Work_num
Productivity_Change_Numeric
Access_to_Mental_Health_Resources_num
Sleep_Quality_num
Degree_of_Remote_Working
Company_Support_for_Remote_Work
Social_Isolation_Rating
Hours_Worked_Per_Week
Number_of_Virtual_Meetings
Physical_Activity_num
Work_Life_Balance_Rating
−3
−2
−1
0
1
SHAP value (impact on model output)
High
Feature value
Low

特徵重要性圖

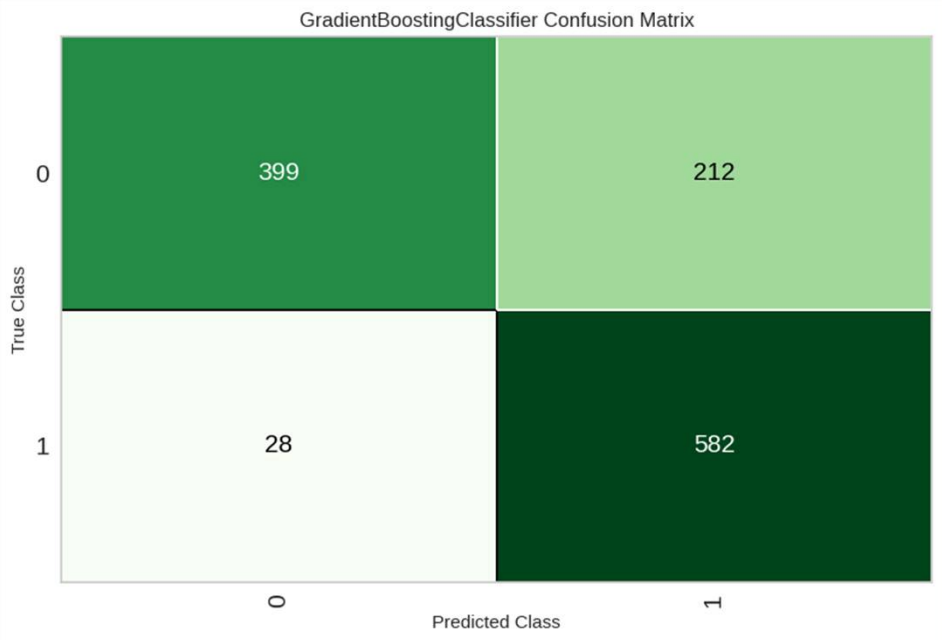
Feature Importance Plot
Hours_Worked_Per_Week
encoded_region
Number_of_Virtual_Meetings
Social_Isolation_Rating
Company_Support_for_Remote_Work
Work_Life_Balance_Rating
Productivity_Change_Numeric
Satisfaction_with_Remote_Work_num
Sleep_Quality_num
Degree_of_Remote_Working
Features
0
100
200
300
400
500
Variable Importance

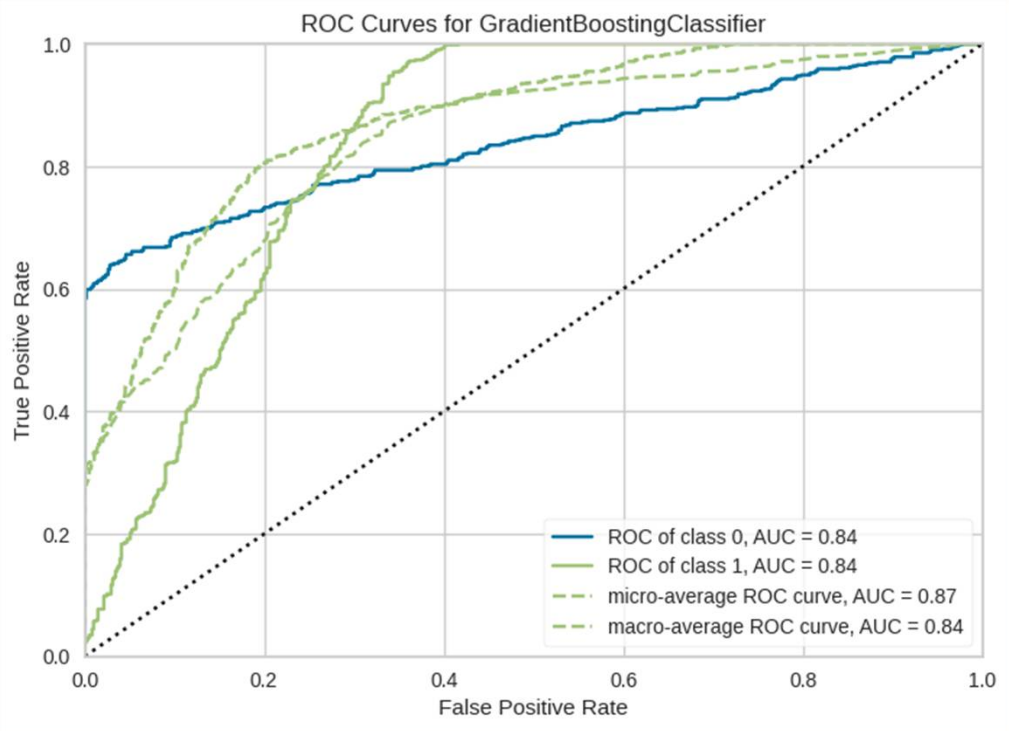
Gradient Boosting Classifier

- 透過組合多個弱學習器來建構一個強學習器的方法
- 以下是此模型在測試集上的表現：

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Light Gradient Boosting Machine	0.7270	0.4906	0.9388	0.7567	0.8380	0.0311	0.0406

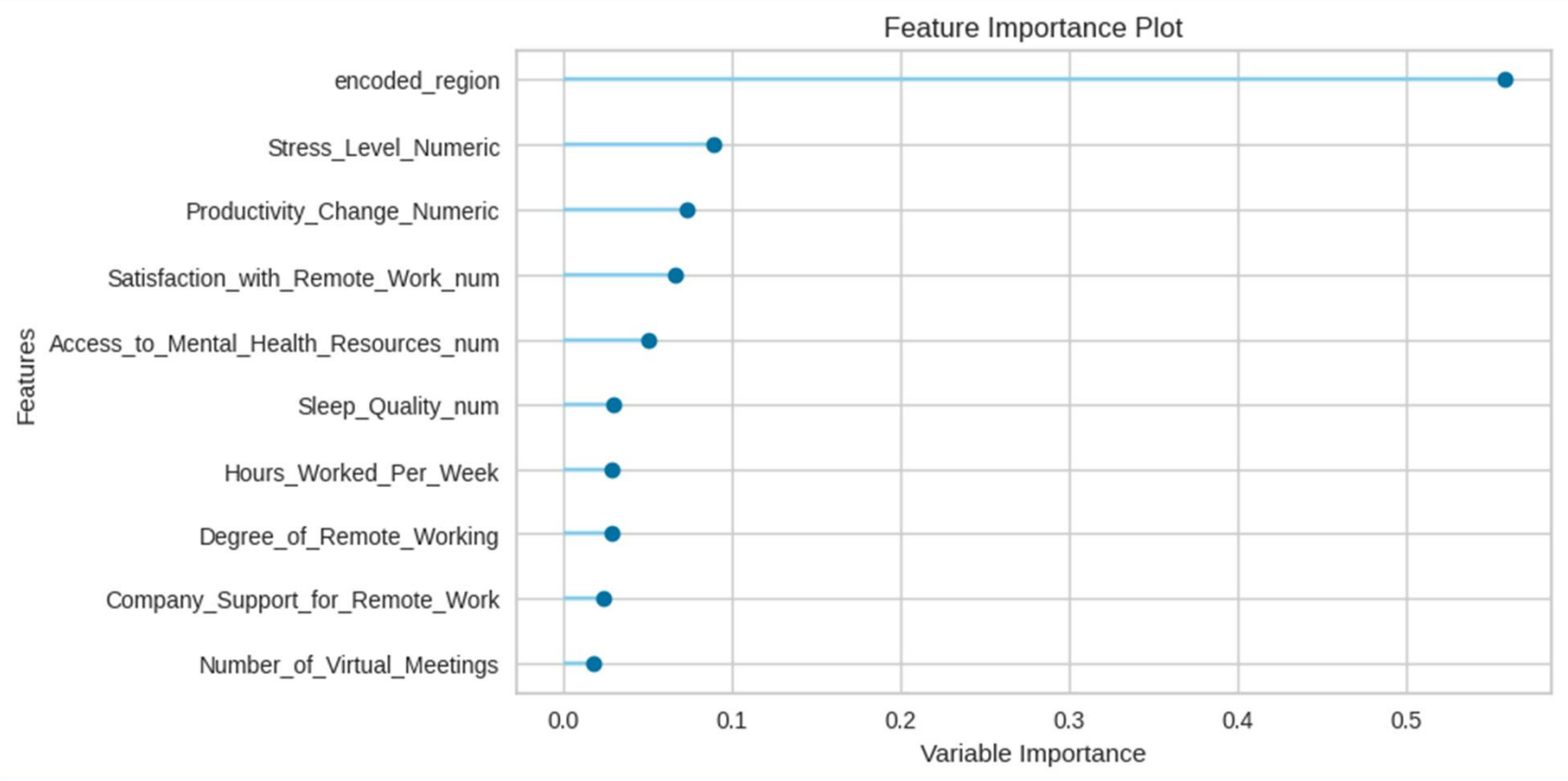
混淆矩陣 & ROC曲線





*在測試集上進行

特徵重要性圖

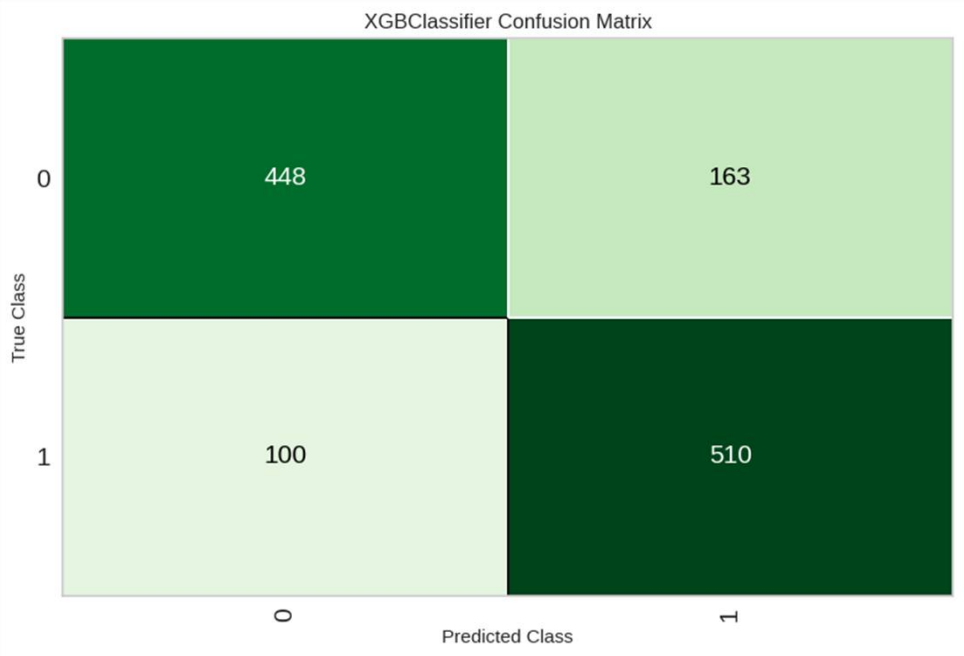

Feature Importance Plot
encoded_region
Stress_Level_Numeric
Productivity_Change_Numeric
Satisfaction_with_Remote_Work_num
Access_to_Mental_Health_Resources_num
Sleep_Quality_num
Hours_Worked_Per_Week
Degree_of_Remote_Working
Company_Support_for_Remote_Work
Number_of_Virtual_Meetings
Features
0.0
0.1
0.2
0.3
0.4
0.5
Variable Importance

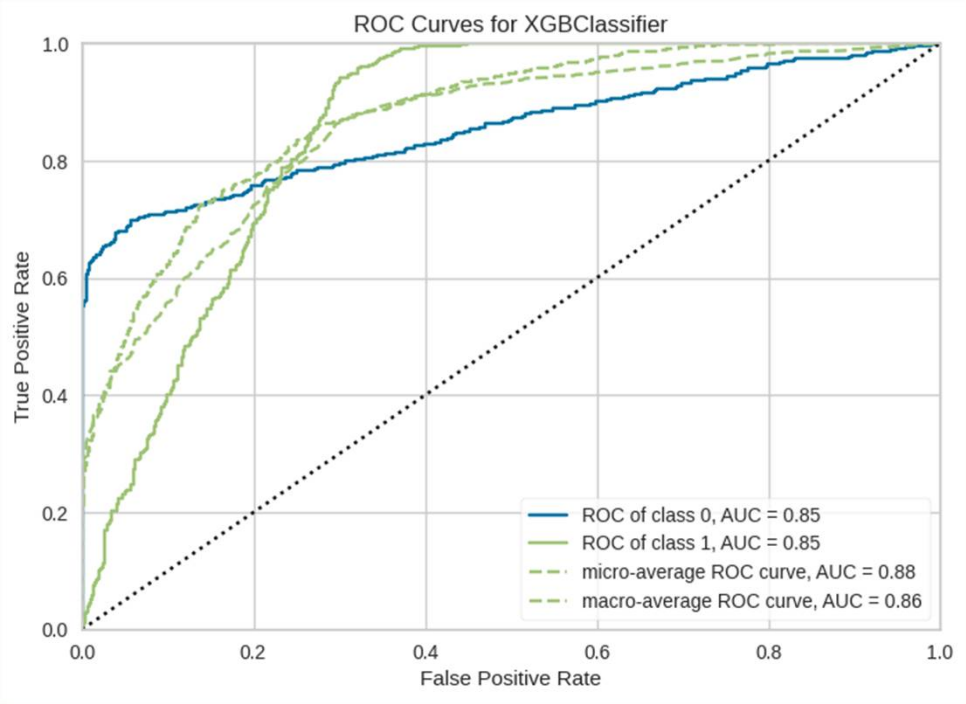
Extreme Gradient Boosting

- 是一種基於梯度提升決策樹 (GBDT) 的機器學習演算法
- 以下是此模型在測試集上的表現：

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Light Gradient Boosting Machine	0.7270	0.4906	0.9388	0.7567	0.8380	0.0311	0.0406

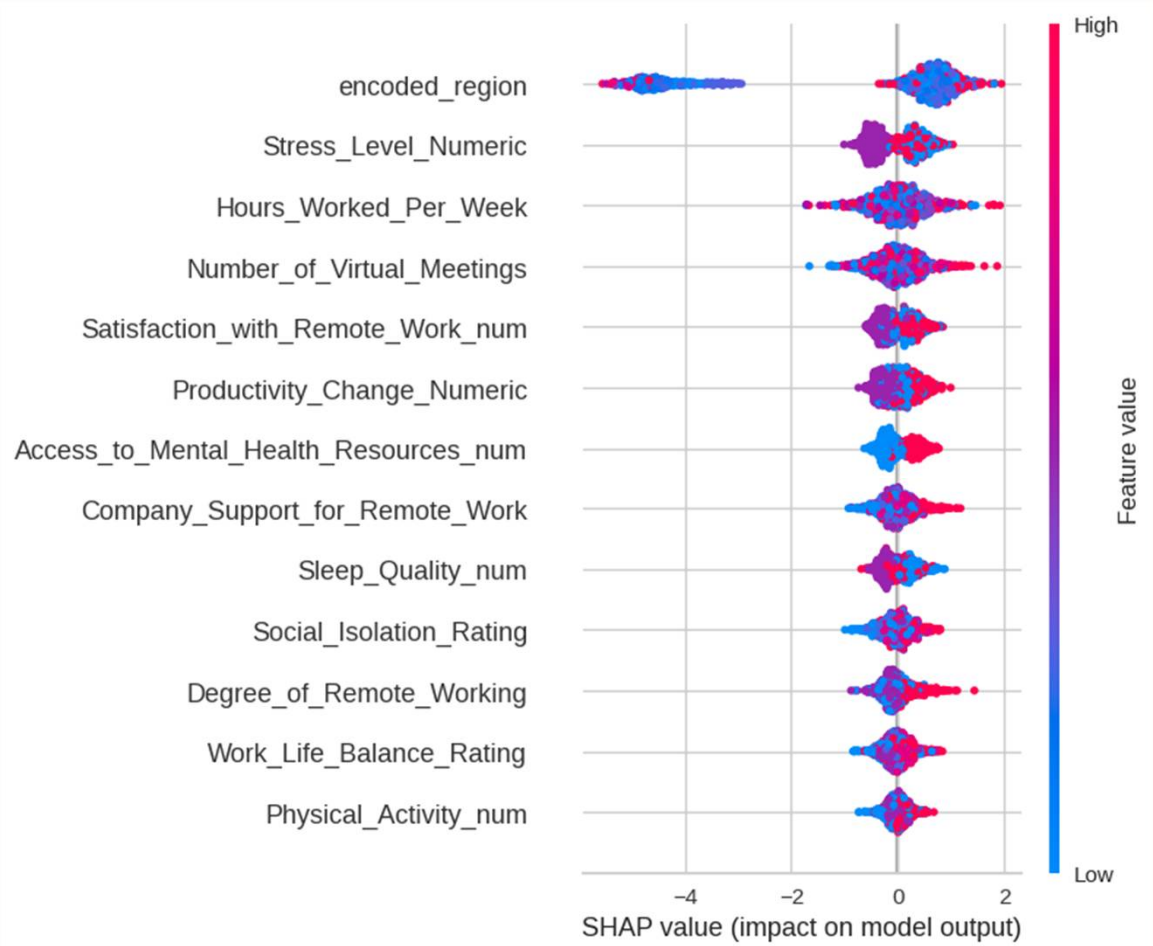
混淆矩陣 & ROC曲線



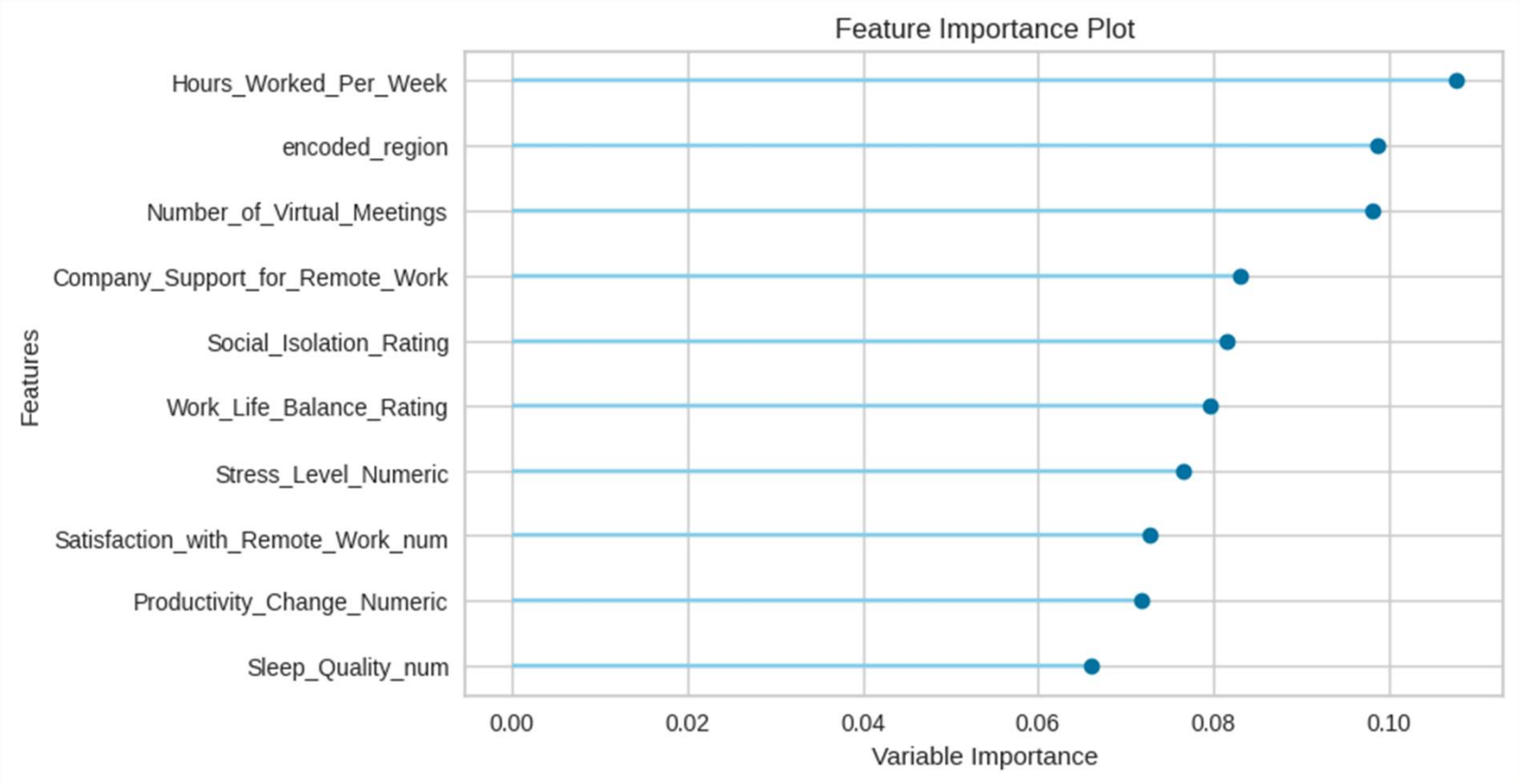


*在測試集上進行

SHAP


encoded_region
Stress_Level_Numeric
Hours_Worked_Per_Week
Number_of_Virtual_Meetings
Satisfaction_with_Remote_Work_num
Productivity_Change_Numeric
Access_to_Mental_Health_Resources_num
Company_Support_for_Remote_Work
Sleep_Quality_num
Social_Isolation_Rating
Degree_of_Remote_Working
Work_Life_Balance_Rating
Physical_Activity_num
−4
−2
0
2
SHAP value (impact on model output)
High
Feature value
Low

特徵重要性圖

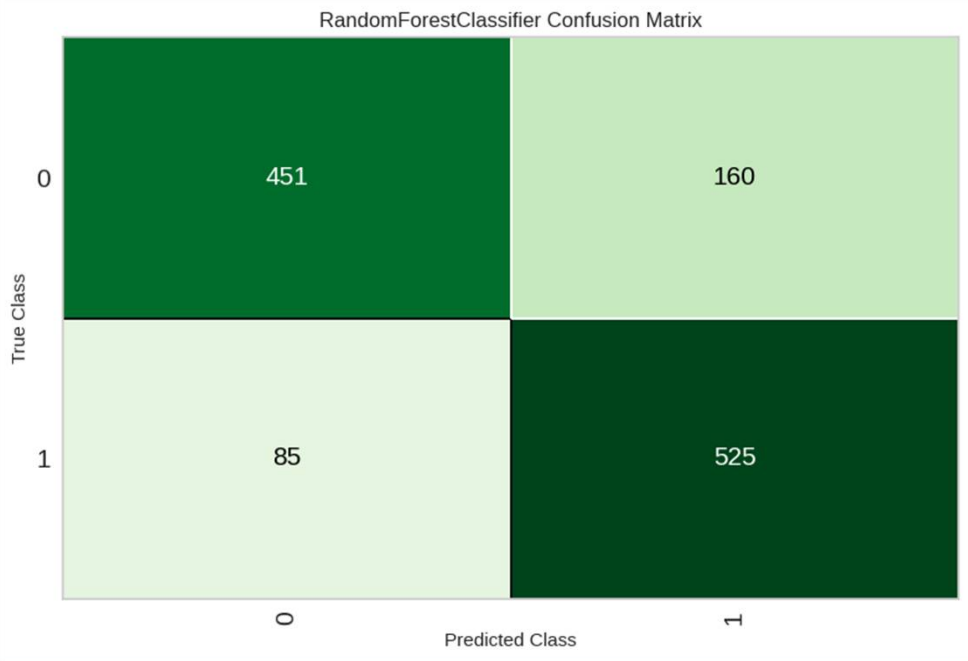

Feature Importance Plot
Hours_Worked_Per_Week
encoded_region
Number_of_Virtual_Meetings
Company_Support_for_Remote_Work
Social_Isolation_Rating
Work_Life_Balance_Rating
Stress_Level_Numeric
Satisfaction_with_Remote_Work_num
Productivity_Change_Numeric
Sleep_Quality_num
Features
0.00
0.02
0.04
0.06
0.08
0.10
Variable Importance

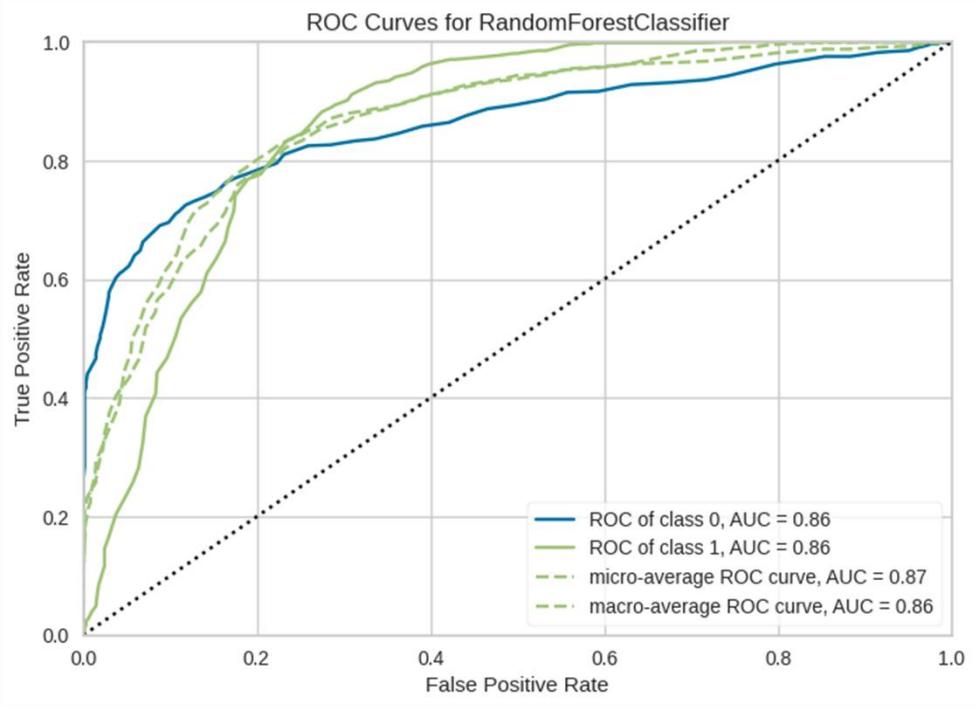
Random Forest Classifier

- 結合了多個決策樹的預測結果來提升整體模型的準確度和穩定性的方法
- 常用於分類問題
- 以下是此模型在測試集上的表現：

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Random Forest Classifier	0.6720	0.5114	0.8604	0.7437	0.7978	-0.0462	-0.0499

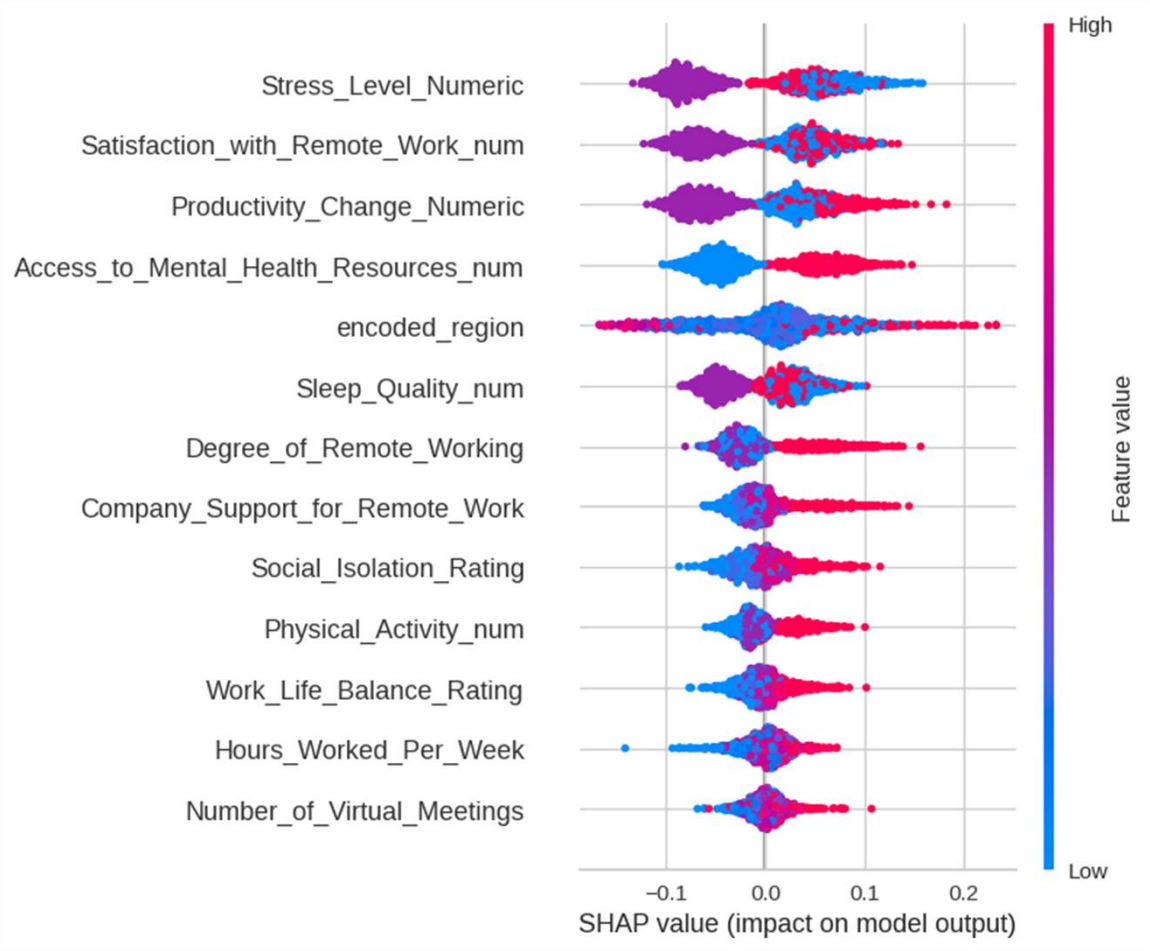
混淆矩陣 & ROC曲線



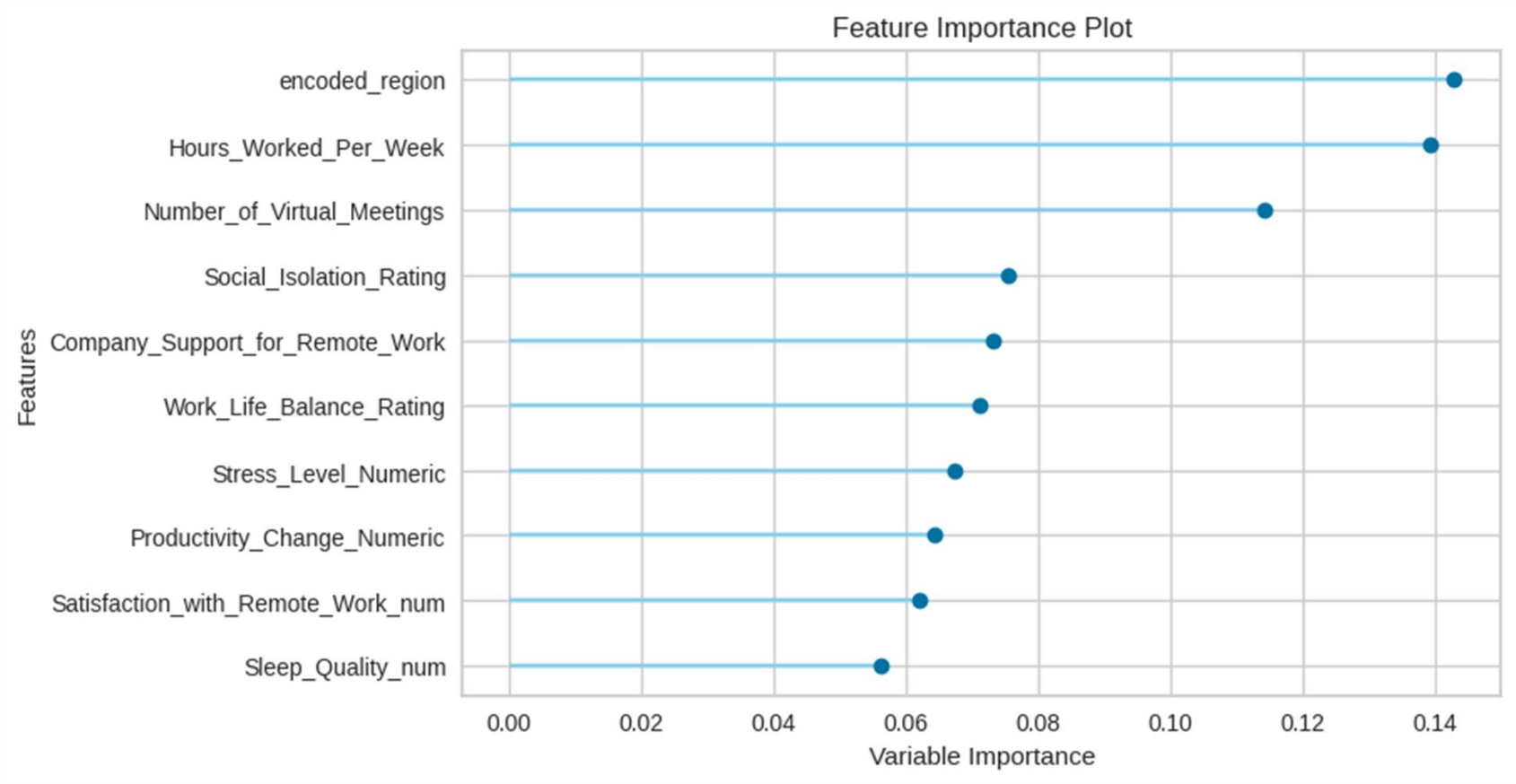


*在測試集上進行

SHAP


Stress_Level_Numeric
Satisfaction_with_Remote_Work_num
Productivity_Change_Numeric
Access_to_Mental_Health_Resources_num
encoded_region
Sleep_Quality_num
Degree_of_Remote_Working
Company_Support_for_Remote_Work
Social_Isolation_Rating
Physical_Activity_num
Work_Life_Balance_Rating
Hours_Worked_Per_Week
Number_of_Virtual_Meetings
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

特徵重要性圖

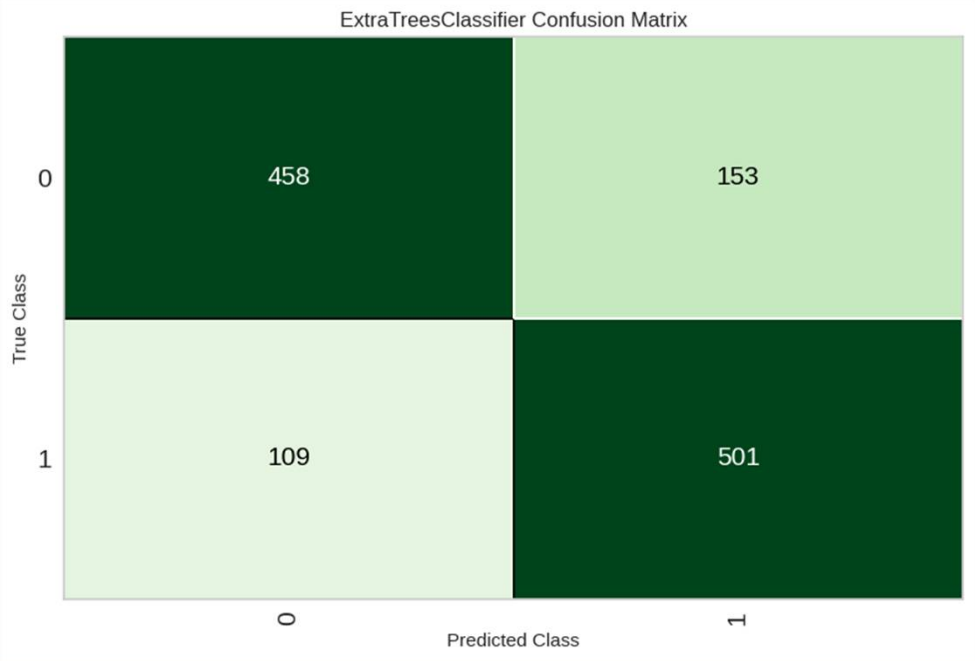

Feature Importance Plot
encoded_region
Hours_Worked_Per_Week
Number_of_Virtual_Meetings
Social_Isolation_Rating
Company_Support_for_Remote_Work
Work_Life_Balance_Rating
Stress_Level_Numeric
Productivity_Change_Numeric
Satisfaction_with_Remote_Work_num
Sleep_Quality_num
Features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Variable Importance

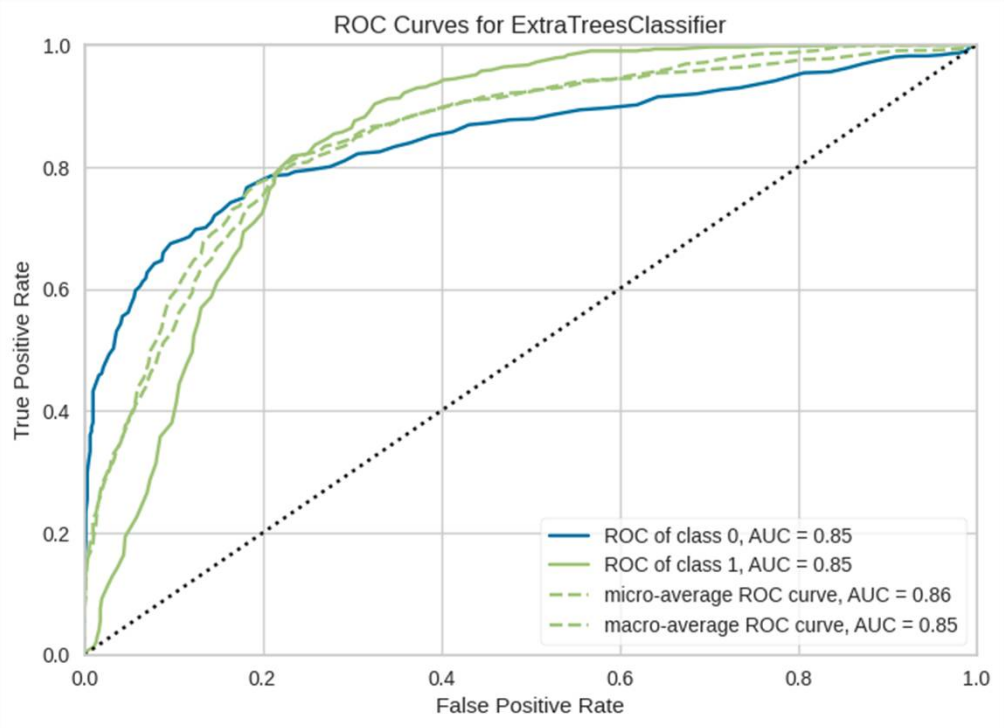
Extreme Trees Classifier

- 是一種集成學習方法，屬於決策樹模型的一種變體
- 與隨機森林 (Random Forest) 非常相似
- 以下是此模型在測試集上的表現：

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Extra Trees Classifier	0.6530	0.4817	0.8271	0.7414	0.7819	-0.0542	-0.0563

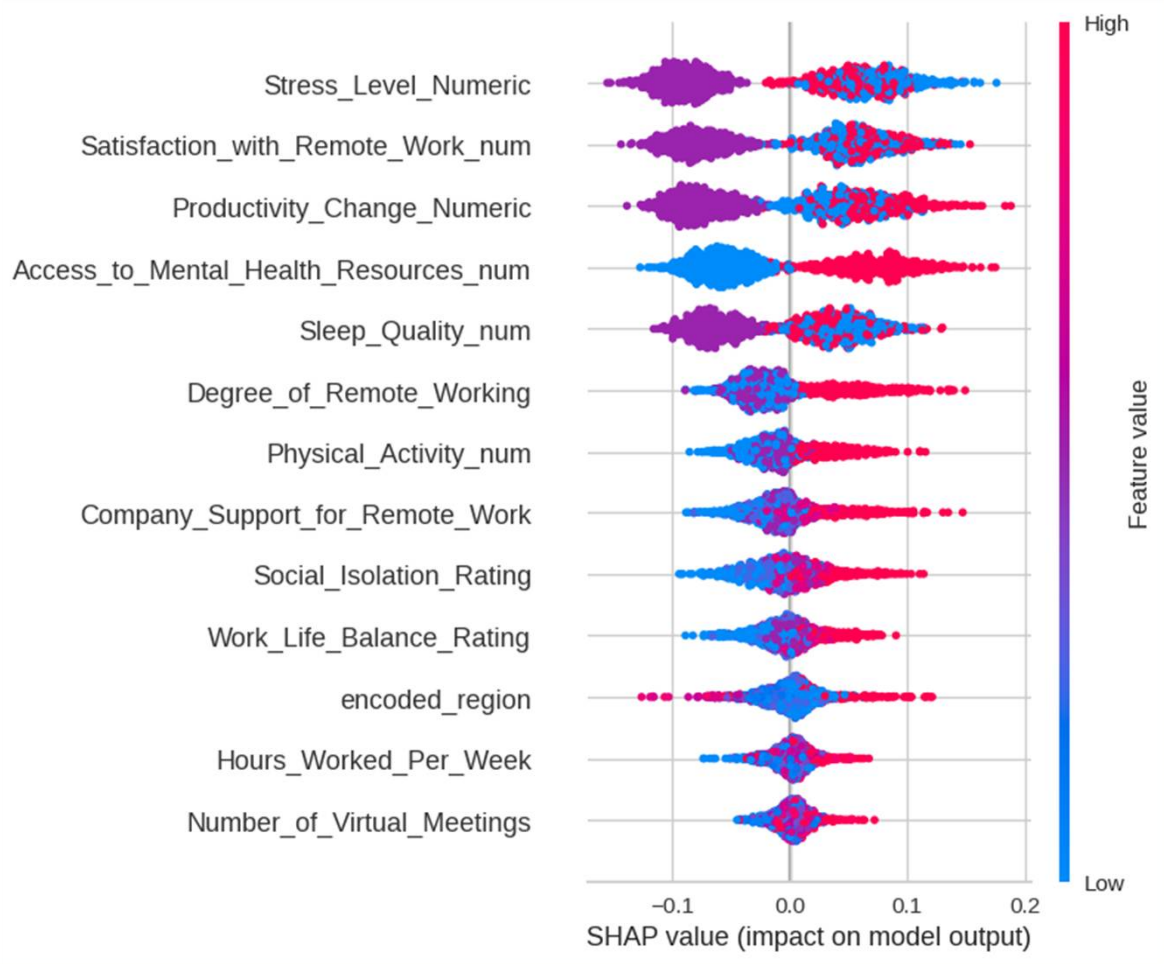
混淆矩陣 & ROC曲線





*在測試集上進行

SHAP


Stress_Level_Numeric
Satisfaction_with_Remote_Work_num
Productivity_Change_Numeric
Access_to_Mental_Health_Resources_num
Sleep_Quality_num
Degree_of_Remote_Working
Physical_Activity_num
Company_Support_for_Remote_Work
Social_Isolation_Rating
Work_Life_Balance_Rating
encoded_region
Hours_Worked_Per_Week
Number_of_Virtual_Meetings
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

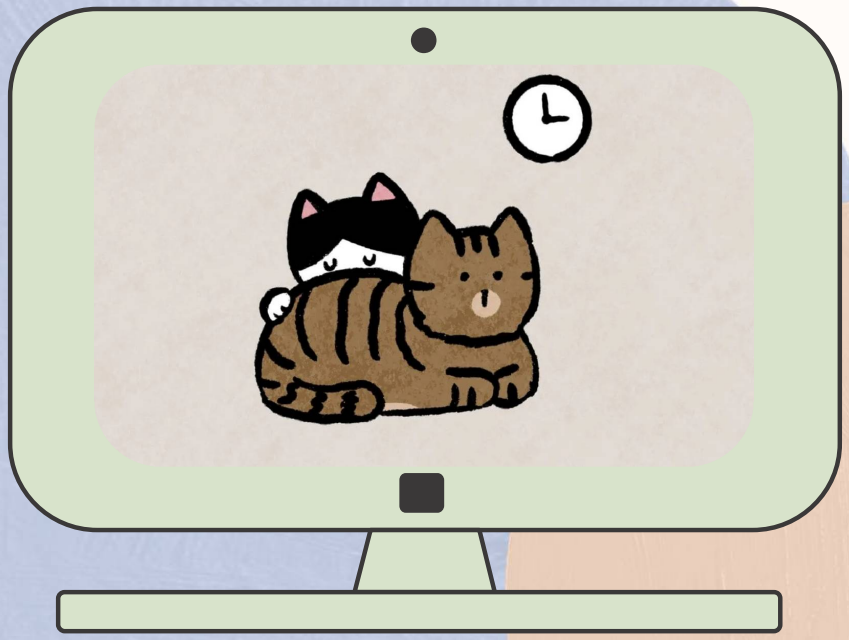
特徵重要性圖

Feature Importance Plot
Hours_Worked_Per_Week
encoded_region
Number_of_Virtual_Meetings
Company_Support_for_Remote_Work
Social_Isolation_Rating
Work_Life_Balance_Rating
Stress_Level_Numeric
Satisfaction_with_Remote_Work_num
Productivity_Change_Numeric
Sleep_Quality_num
Features
0.00
0.02
0.04
0.06
0.08
0.10
Variable Importance

8

結論與未來展望

結論

- CatBoost Classifier 在所有模型中表現最好。
- Region和每周工時以及視訊會議次數，這三個特徵在各個模型中的重要度都是前三名；而睡眠品質都在倒數。
- 在ROC圖中，CatBoost Classifier 與 Light Gradient Boosting Machine的AUC of class 0 和 AUC of class 1 的值都最高，代表模型對兩個類別均有穩定的區分能力。
- 在SHAP圖中，發現每周工時越長，越有可能預測為心理有狀況。
- 希望可以改進特徵工程的部分，看看是否可別的編碼方式可以讓特徵有更加好的表現。

Thanks
for
Listening!