Student Depression

Analyzing Mental Health Trends & Predictors Among Students

Group A（Team 6）

312554010 周鈺祥　312554036 陳胤宏

May 21, 2025
10:00 a.m.

What we'll discuss later

- → Introduction
- → Problem Setting
- → Related Works
- → Proposed Methods
- → Experiment
- → Work distribution chart

Background

" In Maharashtra , student suicides are a severe issue. In 2016, Maharashtra recorded 1,350 student suicides "

Secretary of the Interior

Background

" One of the leading causes of student suicides in India is exam-related failure. More than a 1/4 suicides are linked to academic pressure "

Secretary of the Interior

Motivation

Academic and exam-related stress constitute a significant portion of the pressure faced by Indian students

Research Aims

- **Analyze the relationship between academic pressure, lifestyle habits, economic conditions, and depression among students.**

Challenge

1. The dataset contains a limited number of features, which may restrict model performance even after selection.

2. Correlation analysis on integer-type features showed low relevance to the target, except for academic pressure.

Contribution

- **Most existing studies predominantly utilize Random Forest or Logistic Regression models. In contrast, we are the first to employ a transformer-based model to predict which factors influencing student stress are associated with depression.**

Problem Description

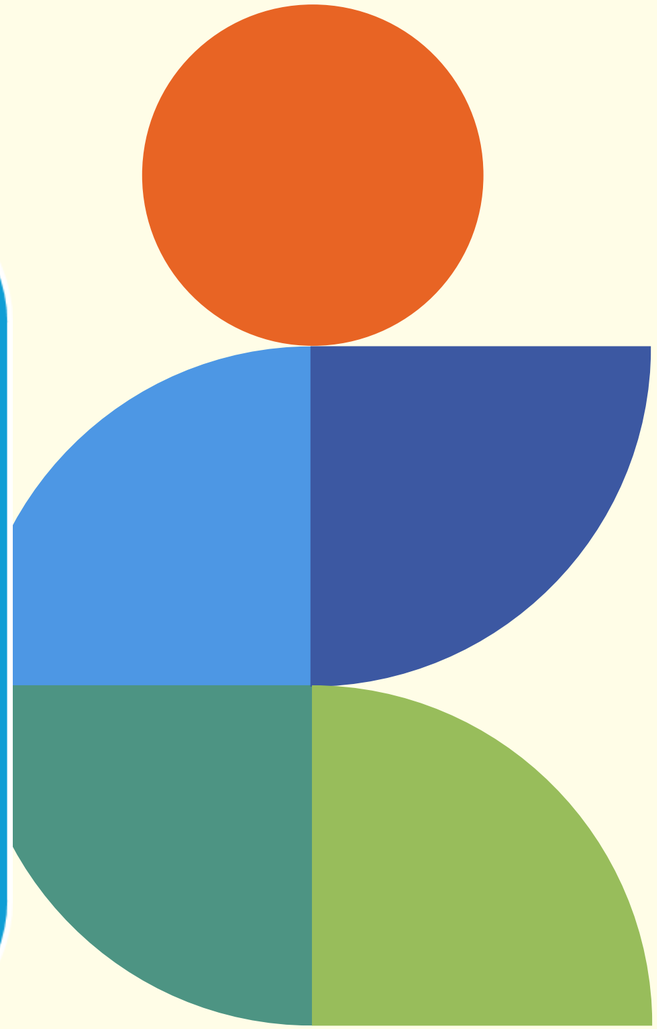
Train a transformer-based model using the dataset to detect whether an individual was experiencing depression at that time

(The dataset does not have any temporal issues)

Input	The dataset includes basic information (e.g., gender, age), academic pressure, daily habits, economic status, and other relevant factors
Process	Analyze the correlation between features and depression, then apply a transformer-based model to predict depression risk
Output	Predicting whether a person has depression. Since depression is strongly linked to suicide risk, early detection and intervention can help reduce suicide rates and prevent severe mental health issues.

Evaluation Metrics

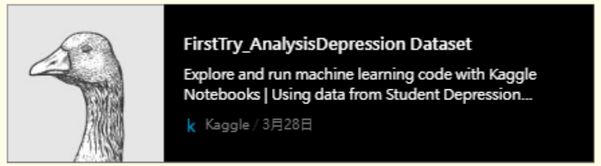
1. **Accuracy**
2. **Precision**
3. **Recall**
4. **F1 Score**



Target Performance

Hybrid Random Forest + XGBoost Hybrid Model

Accuracy	precision	recall	f1-score
0.84	0.84	0.84	0.84



https://www.kaggle.com/code/lukaslieber/firsttry-analysisdepression-dataset

Target Performance

Reach 88% for Accuracy

We believe that since the Transformer has an attention mechanism, we can first use it to filter out highly relevant features before feeding them into the prediction model, which could lead to better performance.

Related Works

Title	**Source**	**Year**	**Author**
Prediction of adolescent depression from prenatal and childhood data from ALSPAC using machine learning	Scientific Reports (https://www.nature.com/articles/s41598-024-72158-9)	2024	Arielle Yoo, Fangzhou Li, Jason Youn, Joanna Guan, Amanda E. Guyer, Camelia E. Hostinar & Ilias Tagkopoulos
Identification of depressive symptoms in adolescents using machine learning combining childhood and adolescence features	BMC Public Health (https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-025-21506-z)	2025	Xinzhu Liu, Rui Cang, Zihe Zhang, Ping Li, Hui Wu, Wei Liu , Shu Li
Academic-related stressors predict depressive symptoms in graduate students: A machine learning study	*Behavioural Brain Research (https://pubmed.ncbi.nlm.nih.gov/39521143/)*	2025	A. F. Bastos, O. Fernandes-Jr, S. P. Liberal, A. J. L. Pires, L. A. Lage, O. Grichtchouk, A. R. Cardoso, L. de Oliveira…
Machine learning models for predicting risk of depression in Korean college students: Identifying family and individual factors	Frontiers in Public Health (https://pubmed.ncbi.nlm.nih.gov/36466485/)	2022	M. Gil, S.-S. Kim, and E. J. Min
Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine	Psychiatry Research (https://pubmed.ncbi.nlm.nih.gov/38690202/)	2021	L. López Steinmetz, J. Godoy, and M. F. Fong

Paper 1

Prediction of Adolescent Depression from Prenatal and Childhood Data from ALSPAC Using Machine Learning

Input

885 features from pregnancy to the child's age of 10

Process

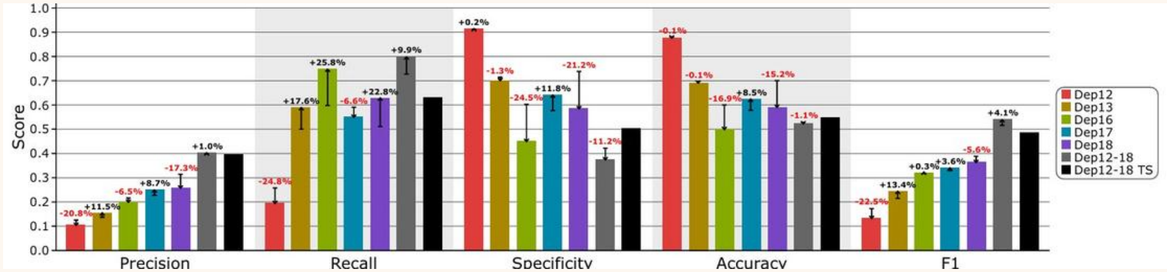
SMOTE
Feature Selection: Pearson correlation coefficient
Recursive Feature Elimination

Output

Predicting whether an individual will experience depression at any time point between the ages of 12 and 18

Algorithm

Classifier : Decision Tree, Naive Bayes, SVM, AdaBoost, Random Forest, MLP (ANN)
Time siries：RNN、LSTM



Paper 2

Identification of depressive symptoms in adolescents using machine learning combining childhood and adolescence features

Input

childhood and adolescence features

Process

Borderline-SMOTE
mRMR
5-fold cross-validation

Output

Whether each adolescent is predicted to have depressive symptoms (Yes/No)

Algorithm

RF、SVM、MLP、SLR

Data sets	Models	Precision	Recall	F1-score	AUC
Training set	Demographic	0.614 ± 0.03	0.613 ± 0.03	0.612 ± 0.03	0.628 ± 0.03
	Child-adolescent_RF	0.806 ± 0.06	0.762 ± 0.09	0.750 ± 0.09	0.884 ± 0.02
	Child-adolescent_SLR	0.838 ± 0.01	0.834 ± 0.00	0.833 ± 0.00	0.879 ± 0.01
	Child-adolescent_SVM	0.857 ± 0.01	0.852 ± 0.01	0.851 ± 0.01	0.885 ± 0.01
	Child-adolescent_MLP	0.855 ± 0.02	0.851 ± 0.01	0.850 ± 0.01	0.907 ± 0.02
	Combined_RF	0.795 ± 0.04	0.792 ± 0.04	0.792 ± 0.04	0.861 ± 0.01
	Combined_SLR	0.834 ± 0.01	0.830 ± 0.01	0.829 ± 0.01	0.882 ± 0.01
	Combined_SVM	0.841 ± 0.02	0.837 ± 0.02	0.836 ± 0.02	0.886 ± 0.01
	Combined_MLP	0.850 ± 0.03	0.847 ± 0.03	0.847 ± 0.03	0.901 ± 0.02
Test set	Demographic	0.798	0.495	0.591	0.530
	Child-adolescent_RF	0.872	0.763	0.803	0.835
	Child-adolescent_SLR	0.913	0.845	0.868	0.879
	Child-adolescent_SVM	0.920	0.814	0.846	0.876
	Child-adolescent_MLP	0.889	0.784	0.820	0.866
	Combined_RF	0.879	0.804	0.832	0.840
	Combined_SLR	0.900	0.845	0.865	0.864
	Combined_SVM	0.907	0.814	0.845	0.876
	Combined_MLP	0.875	0.784	0.817	0.857

Paper 3

Academic-related stressors predict depressive symptoms in graduate students: A machine learning study

Input

10 academic stressor scores
PHQ-9 depression scores
Demographics

Algorithm

ε-SVR

Process

ε-SVR model with k-fold cross-validation
Feature importance analysis

Regression Models	r	*p-value*	R^2	*p-value*	NMSE	*p-value*
ε-SVR (k = 5)	0.44	*0.001*	0.21	*0.001*	0.84	*0.001*
ε-SVR (k = 2)	0.47	*0.001*	0.22	*0.001*	0.78	*0.001*

Output

Predicted PHQ-9 scores

Paper 4

Machine learning models for predicting risk of depression in Korean college students: Identifying family and individual factors

Input

CES-D depression scores, Big Five personality traits, parental health conditions, attachment styles, family cohesion, and demographic data

Algorithm

LASSO、SVM、RF

Process

Labeling students based on CES-D scores (≥13 = at-risk)
Feature processing and standardization
Feature selection via LASSO and feature importance via Gini index (RF)

Output

Predicted depression risk for each student

	Accuracy	Precision	Sensitivity	Specificity	F1 score	AUC
Sparse logistic	0.7843	0.7143	0.7500	0.8065	0.7317	0.9032
SVM	0.8039	0.7500	0.7500	0.8387	0.7500	0.7944
Random forest	0.8627	0.8059	0.8500	0.8710	0.8274	0.8605

Paper 5

Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine

Input

Sociodemographic and lifestyle data from 329 university students in Lebanon two university

Process

Missing values imputed using mode
Feature importance analyzed using Random Forest

Output

Depression
Anxiety
Stress

Algorithm

LR、RF、SVM、KNN、AdaBoost、XGBoost、Naïve Bayes、MLP

Model	AUC (%)	Sensitivity (%)	Precision (%)	*F1*-score (%)
MLP				
Depression	73.90	74.28	63.41	68.42
Anxiety	72.60	73.68	63.63	68.29
Stress	70.30	66.66	47.05	55.17
Logistic regression				
Depression	74.12	55.00	78.00	64.00
Anxiety	74.89	74.00	75.00	74.00
Stress	66.51	7.00	100.00	13.00
AdaBoost				
Depression	76.25	65.00	72.00	68.00
Anxiety	74.89	100.00	53.00	69.00
Stress	72.96	33.00	69.00	45.00
Random forest				
Depression	78.27	60.00	75.00	67.00
Anxiety	69.93	62.00	70.00	66.00
Stress	72.42	12.00	75.00	19.00
XGBoost				
Depression	75.55	51.00	79.00	68.00
Anxiety	67.67	71.00	65.00	68.00
Stress	66.87	30.00	72.00	42.00
SVM				
Depression	74.36	67.00	65.00	66.00
Anxiety	74.94	72.00	72.00	72.00
Stress	72.37	3.00	100.00	7.00
Naïve Bayes				
Depression	74.12	62.00	71.00	65.00
Anxiety	76.37	77.00	75.00	76.00
Stress	63.36	37.00	47.00	41.00
KNN				
Depression	66.63	35.00	66.00	46.00
Anxiety	61.05	53.00	57.00	62.00
Stress	63.84	12.00	60.00	18.00

Abbreviations: AUC, area under curve; *F1*-score, harmonic mean between precision and recall.

Summary

Title	Sample size	Data source	Data type	Target
Prediction of adolescent depression …	15,645	ALSPAC	Numeric: 5-600 Binary (0/1): 2-300	Depression
Identification of depressive …	505	Peking University Institute of Social Science Survey	Numeric: 18 Binary (0/1): 13 Object / Categorical: 8	Depression
Academic-related stressors …	172	Questionnaire survey (Non-public)	Numeric: 11 Binary (0/1): 1 Object / Categorical: 3	PHQ-9 depression score
Machine learning models for ..	171 families (513 individuals)	Questionnaire survey (Non-public)	Numeric: ~5 Binary (0/1): ~10 Object / Categorical: ~20	Depression
Machine learning models predict …	329	Online questionnaire survey (Non-public)	Numeric: 1 Binary (0/1): 6 Object / Categorical: 11	Depression, Anxiety, Stress

Summary

Title	Insights/Pitfalls	Application design result
Prediction of adolescent depression …	Insights:Recursive Feature Elimination (RFE) effectively reduces feature dimensionality. Pitfalls:Severe issues with data imbalance and missing values.	The number of features was reduced from 266 to only 14 , improving the model's interpretability and deployment feasibility
Identification of depressive …	Insights:Cross-period features can effectively improve prediction. Pitfalls:Small sample size and only self-administered questionnaires were used, which lacked clinical validation.	Why use mRMR? Emphasis on interpretability and avoiding collinearity
Academic-related stressors …	Insights:ε-SVR provides a “weight” for each stress factor, helping to understand which factors have the greatest impact. Pitfalls:The sample was not randomly selected, which may be biased, and only self-administered questionnaires were used, which lacked clinical validation.	Why choose ε-SVR? Because the sample size is small and SVR is stable in small samples
Machine learning models for ..	Insights:Family-related factors play a significant role in student mental health, especially in Korean culture with strong family interdependence Pitfalls:The data normalization method is not described.	Using SLR to provide interpretability and RF to provide high accuracy complements each other.
Machine learning models predict …	Insights:The data comes from two very different universities, which helps improve generalization ability. Pitfalls:Small sample size , online sampling restrictions and subjective bias.	Why only use a questionnaire? Small amount of data, high algorithm interpretability

Conclusion

Referenced Methods

1.SMOTE
2.5-fold cross-validation
3.Feature importance via Gini index (RF)

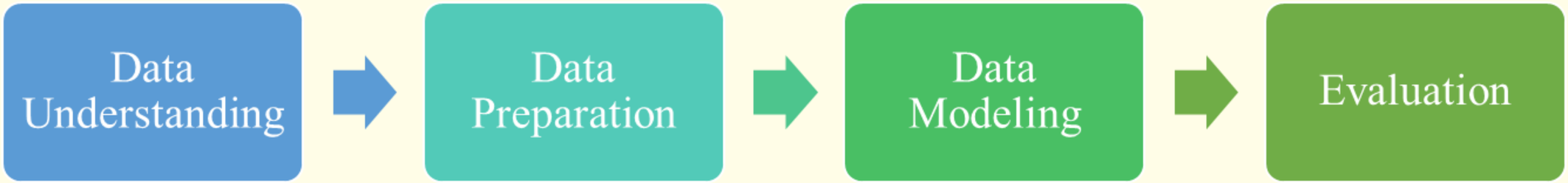
Conclusion

Since the ratio of the number of people with depression to the number of people without depression in the dataset we used is about 4:6, we used SMOTE to smooth the data.

We also added k-fold cross-validation to avoid overfitting the training data.

We also used RF Gini index to check the importance of features and decide whether to add them to the model and how to add new features to improve accuracy.

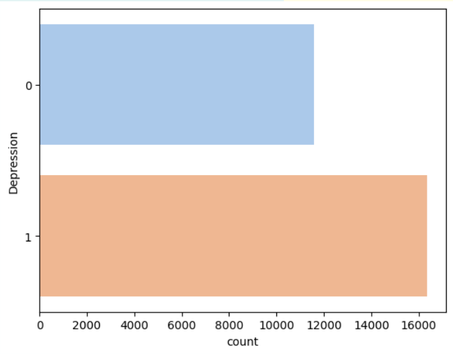
Proposed Methods



Data Understanding

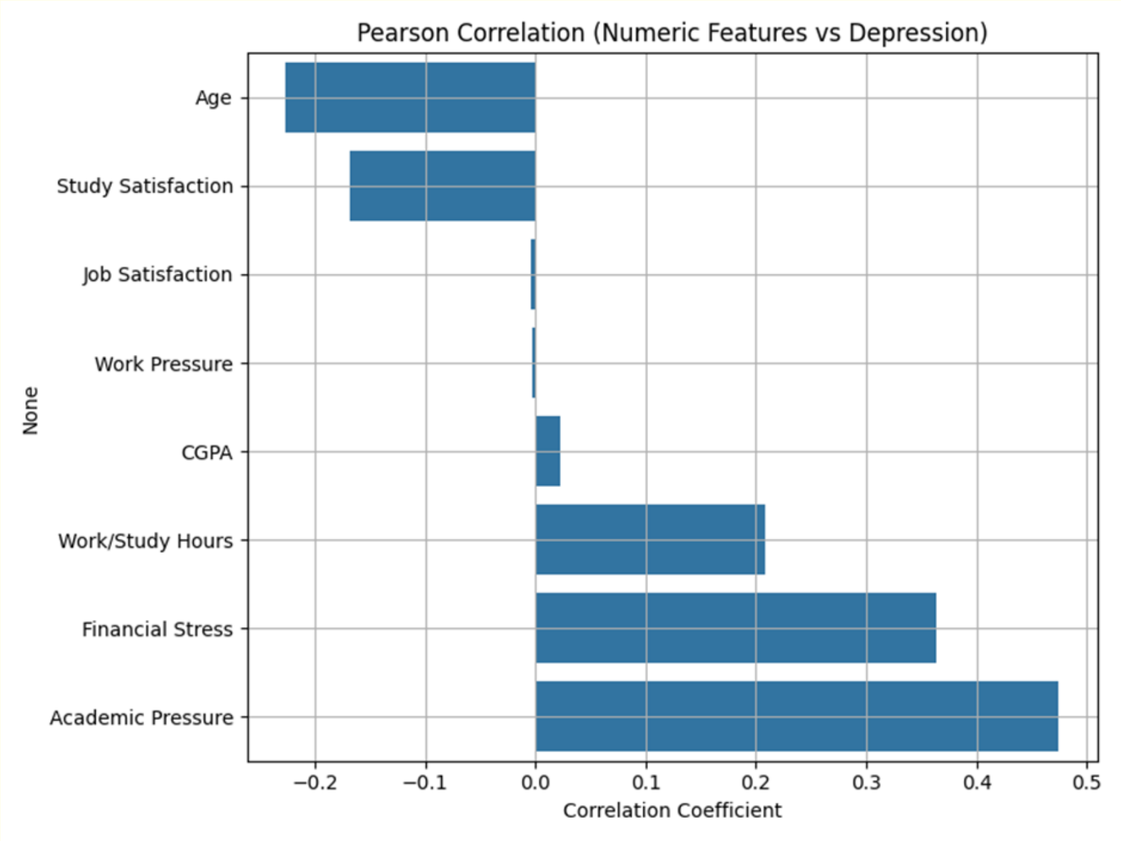
本研究使用的學生心理健康資料集包含 27,901 筆樣本，共計 17 個特徵與 1 個二元標籤（Depression：1 表示有憂鬱傾向，0 表示無）。
特徵類型包括：
數值型特徵：如 CGPA, Age, Academic Pressure, Work/Study Hours ...

類別型特徵：如 Gender, Sleep Duration, Degree, Dietary Habits, City ...

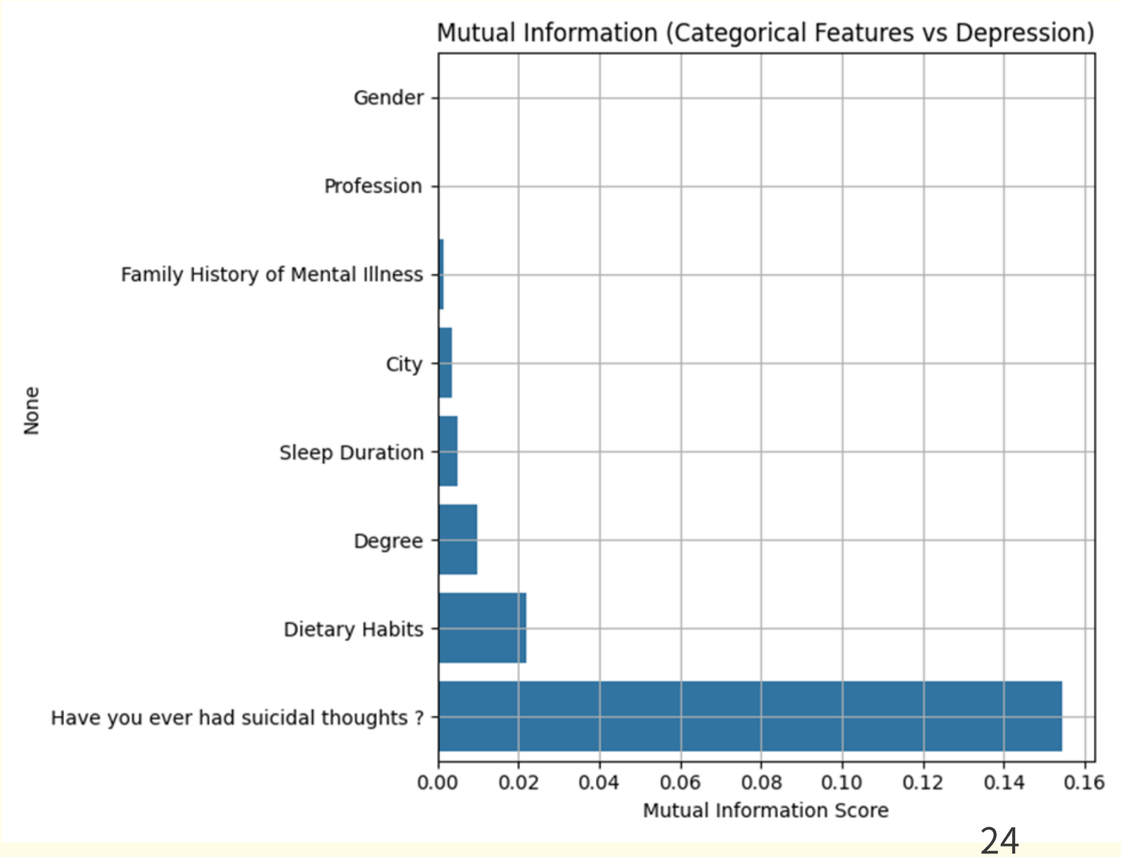


Data Understanding

大多數數值型特徵與 Depression 標籤的相關性偏低



以互資訊評估類別特徵與標籤之間的資訊量



Data Preparation

(1) 類別資料處理
將 **Sleep Duration, Suicidal Thoughts, Family History of Mental Illness** 等欄位進行 **Label Encoding** 編碼。

將出現次數過少的類別（如城市、科系等）合併為 **"Other"** 類別，以避免類別稀疏導致模型過擬合。

填補類別遺漏值

(2) 數值資料處理
將所有數值型欄位（如 **CGPA, Academic Pressure** 等）轉為浮點數格式，並以欄位平均值進行缺失值填補。

使用 **StandardScaler** 對所有數值特徵進行標準化處理。

Data Preparation

特徵相關性熱圖放在補充資料

(3) 特徵工程（Feature Engineering）

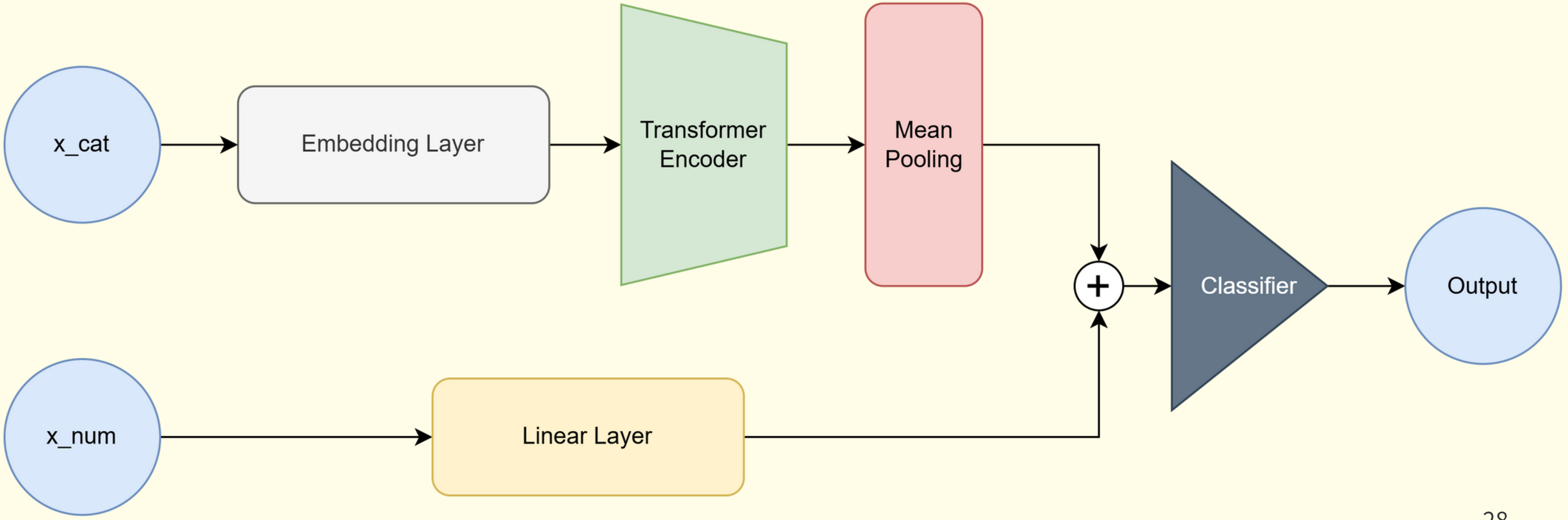
衍生特徵	說明
StressIndex	Academic + Financial + Work 三項壓力指標總和，統整壓力源
BalanceScore	Study Satisfaction + Sleep Duration – Work/Study Hours，衡量生活平衡
CGPA_per_Hour	*學業效率指標，反映在高工時下是否維持良好成績*
CGPA_x_AcadPressure	成績與壓力的交互項，揭露高壓下成績表現差異
CGPA_x_StudySat, Age_x_WorkHours	建構非線性與個人背景交互的判別依據
CGPA_squared, Age_log	對原始欄位進行指數轉換，捕捉非線性關係

Data Preparation

(4) 類別不平衡處理

原始資料中，**Class 1**（憂鬱樣本）相對較多，但仍可能產生學習偏誤。為強化模型泛化能力，採用 **SMOTE**（**Synthetic Minority Oversampling Technique**）平衡訓練集中的樣本分布。

Data Modeling



Evaluation

使用 Binary Cross Entropy（BCE）作為主損失函數，處理二元分類問題。

搭配 AdamW 優化器（學習率為 1e-4，weight decay 為 1e-5）

模型訓練時使用交叉驗證與 Early Stopping 避免過擬合

F1-score 作為主要的效能指標，並搭配 Accuracy 與 Precision、Recall做為輔助分析

通過 Precision-Recall Curve 動態選擇最佳閾值

Data Description

Date Source

Kaggle

Quantity

27901 rows 2.9MB

Columns

18

https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset

Data Description (Cont.)

Name	Type	Domain	Description
ID	Numerical		27901 rows
Gender	Categorical	2(cat)	Male、Female
Age	Numerical		Around 18-59 years old
City	Categorical	52(cat)	Cities in India
Profession	Categorical	14(cat)	Occupation **(Student has 27870 rows)**

Data Description (Cont.)

Name	Type	Domain	Description
Academic Pressure	Categorical	5(cat)	1 represents Low, 5 represents High
Work Pressure	Categorical	5(cat)	1 represents Low, 5 represents High
CGPA	Numerical	0-10	Cumulative Grade Point Average, which is a quantitative measure of the student's academic performance.
Study Satisfaction	Categorical	5(cat)	1 represents High, 5 represents Low
Job Satisfaction	Categorical	5(cat)	1 represents High, 5 represents Low

Data Description (Cont.)

Name	Type	Domain	Description
Sleep Duration	Categorical	5(cat)	less than 5 hours, 5-6 hours, 7-8 hours, more than 8 hours, others
Dietary Habits	Categorical	4(cat)	Healthy, Moderate, Unhealthy, Others
Degree	Categorical	4(cat)	Bachelor, Master, PhD, Others
Suicidal Thoughts	Categorical	2(cat)	Yes/No
Work/Study Hours	Numerical	0-12	Study duration

Data Description (Cont.)

Name	Type	Domain	Description
Financial Stress	Categorical	5(cat)	1 represents Low, 5 represents High
Family History of Mental Illness	Categorical	2(cat)	Yes/No
Depression	Categorical	2(cat)	Yes/No

Evaluation Strategy

資料分割方式

80% 作為訓練 / 驗證資料：供模型訓練與交叉驗證使用

20% 作為獨立測試集（Hold-out Test Set）：不參與任何訓練與參數調整，用於最終效能評估

5-Fold Cross-Validation

每次使用其中一個 fold 做為驗證，其餘 4 folds 作為訓練

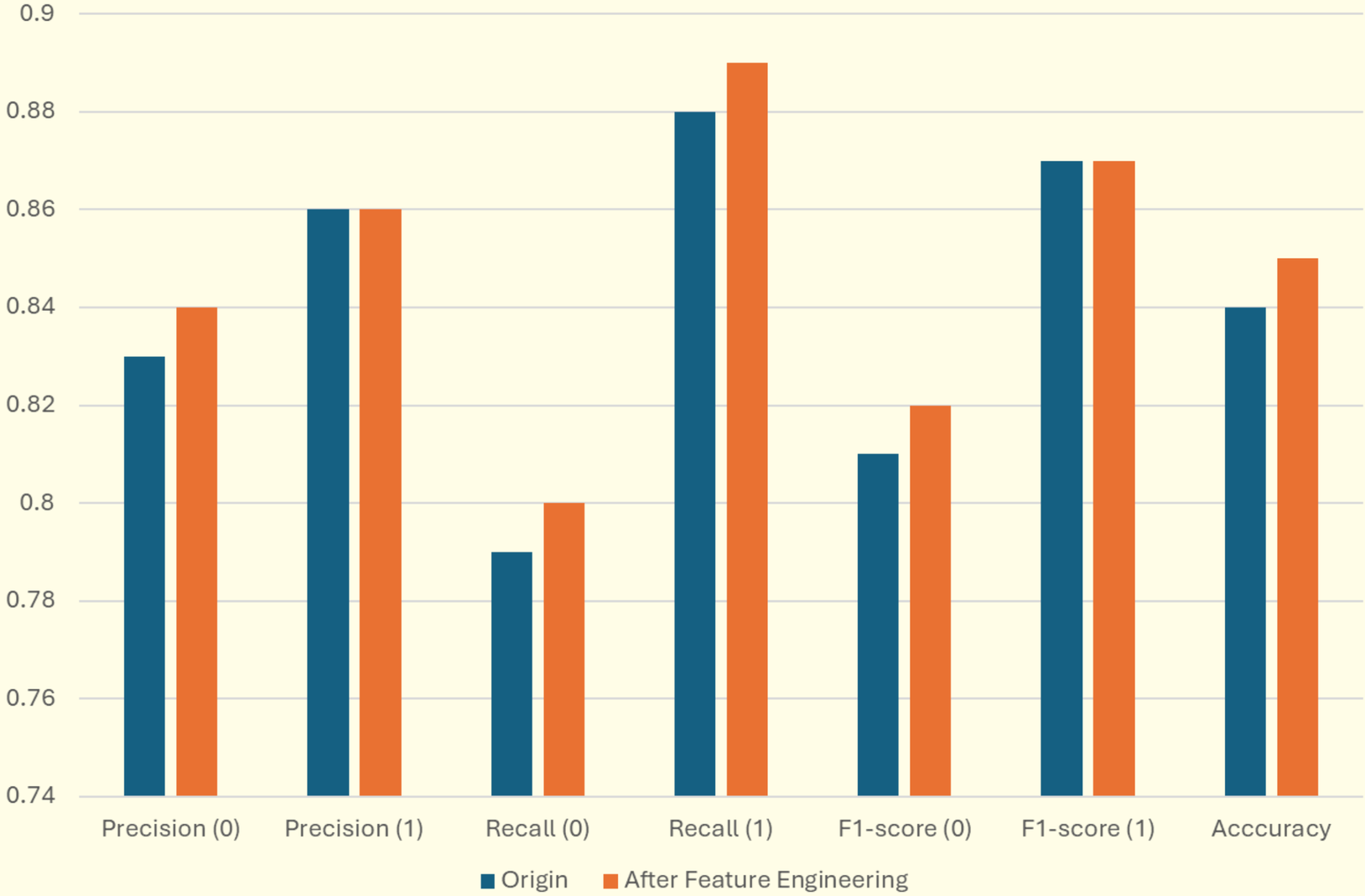
Early Stopping 與 Best Threshold Selection

若驗證集 F1-score 在連續數個 epoch 未提升，則提早終止訓練並保留最佳模型參數
透過 **Precision-Recall 曲線** 計算不同閾值下的 F1-score，動態選擇最佳 decision threshold，而非使用固定 0.5，進一步提升模型分類效能與 recall 能力

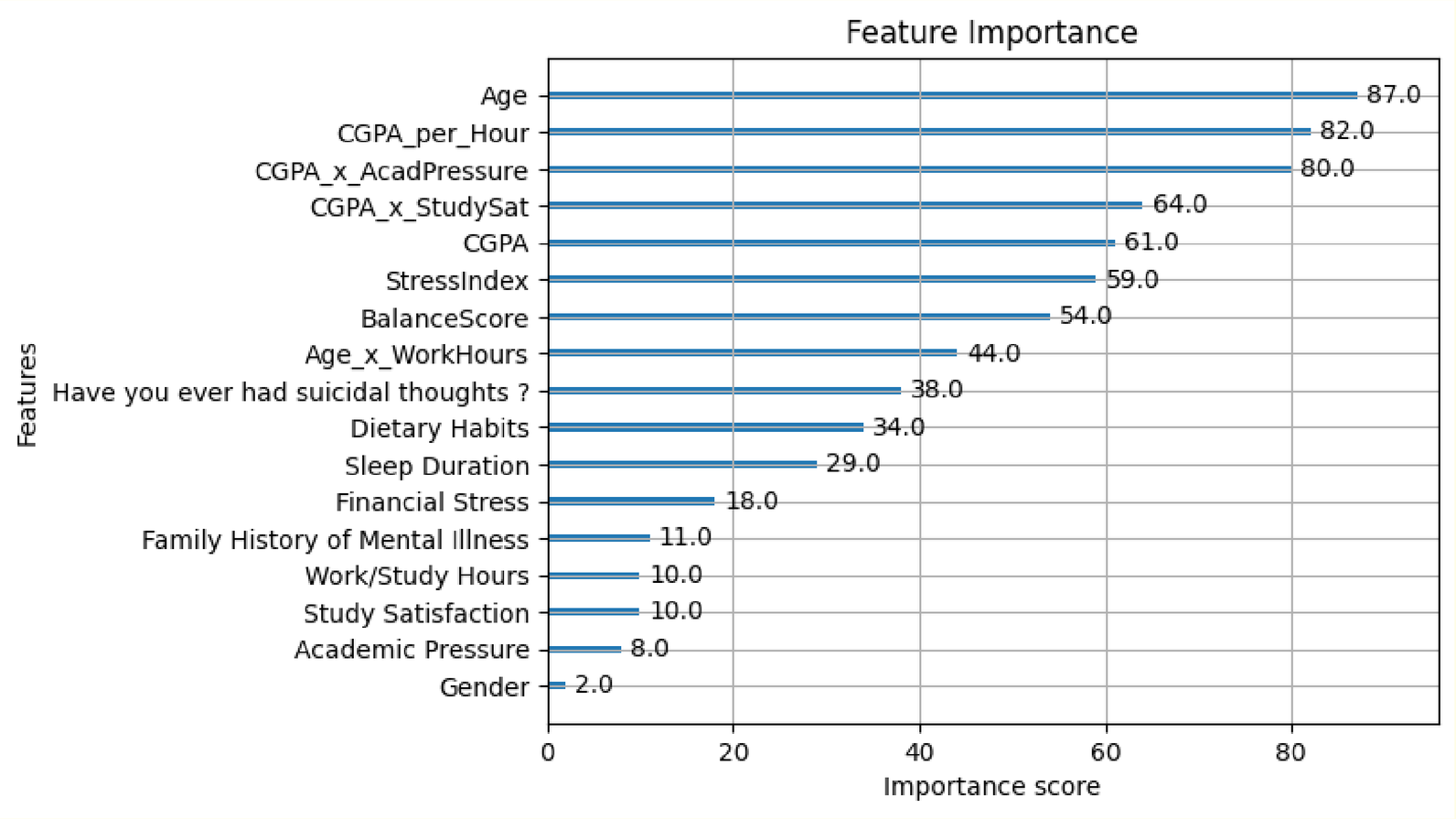
最終測試集驗證

模型訓練與交叉驗證完成後，最終以其中一個最佳模型（從 K-Fold 中挑選）於事先保留的 Hold-out Test Set 上進行推論，並回報以下效能指標：
Accuracy、F1-score、Precision / Recall（特別關注憂鬱類別）

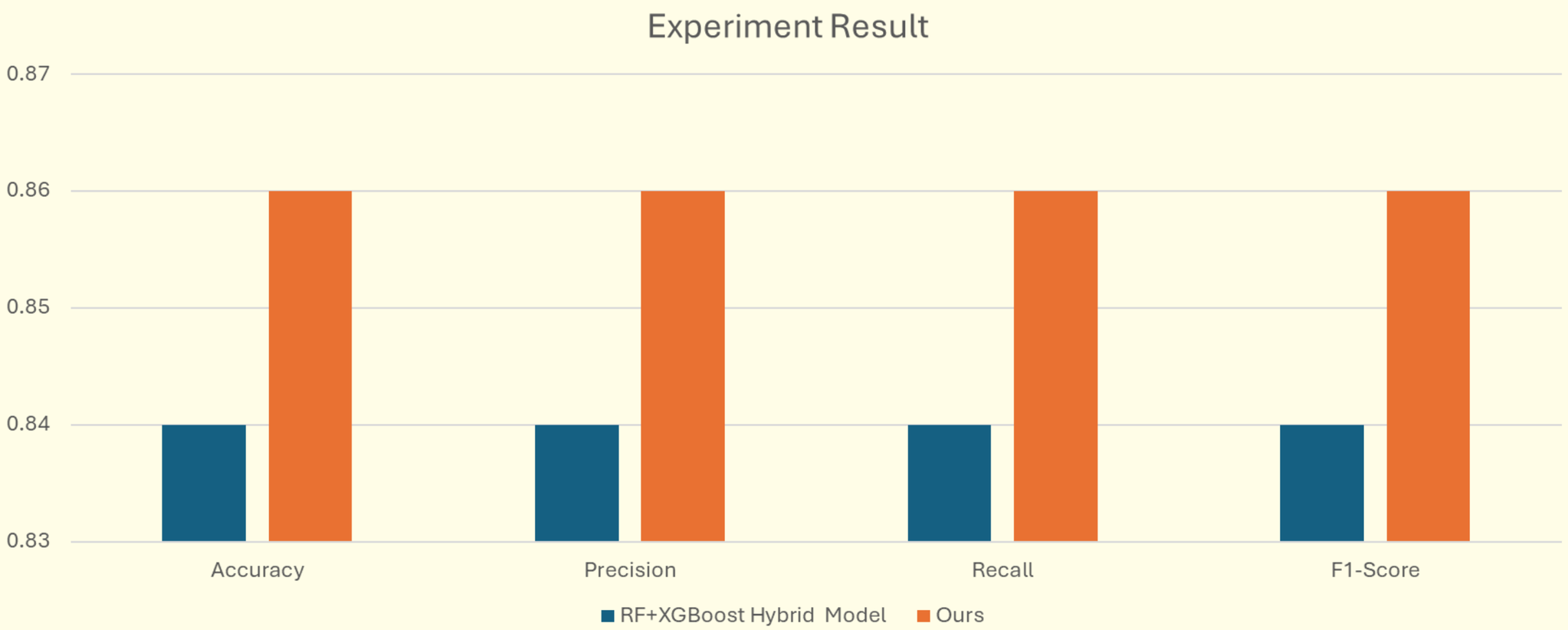
Experimental Results



Experimental Results



Experimental Results



Insights or Thoughts

Transformer
架構對混合型資料有效

模型透過類別注意力結構與 Transformer 對類別特徵進行表徵學習，搭配數值特徵的嵌入與融合，能穩定學習複雜資料關係，尤其適合像學生調查資料這種類別數值混合型資料。

特徵工程明顯提升模型效能

所設計的乘積項（如 CGPA × Academic Pressure）、壓力指標（StressIndex）與比例項（CGPA per Hour）等複合特徵，有助於模型更精準地掌握心理壓力與學業表現之間的非線性關係。

Test Set 表現略高於交叉驗證平均是合理現象

模型在 Test Set 上的 F1-score（0.88）略優於 K-Fold 平均（0.87），可能因 Test Set 噪音較少或與訓練分布較接近所致，此現象合理，亦顯示模型未發生過擬合。

Environment

Environment Setting	
OS	Linux Ubuntu 22.04
Programming Language	Python
Tools	Pytorch, Jupyter Notebook
GPU	4080s
Librarys	numpy, pandas, seaborn, matplotlib, torch

Conclusion

我們提出一套基於 Transformer 的學生憂鬱預測模型，結合精緻的特徵工程策略與 SMOTE 類別平衡技術，並應用動態閾值調整與五折交叉驗證以提升模型的穩定性與泛化能力。
實驗結果顯示，所提出模型於 5-Fold Cross-Validation 中平均 F1-score 達 0.87，並於保留測試集上取得 F1-score 0.88 與 Recall 0.91（Class 1）的優異表現。這代表模型能有效識別具潛在憂鬱風險的學生，具備實務應用潛力，特別適合用於早期介入、心理諮詢篩選或風險預警系統。

Limitation

資料來源與樣本代表性

本研究資料來自特定問卷樣本，可能受限於填答族群的背景、區域或文化差異，未必能直接套用於其他地區或學校的學生族群。

資料型態限制

所用特徵多為靜態、自我報告型變數，尚未納入行為數據（如手機使用模式、課業變化等），導致部分潛在風險難以察覺。

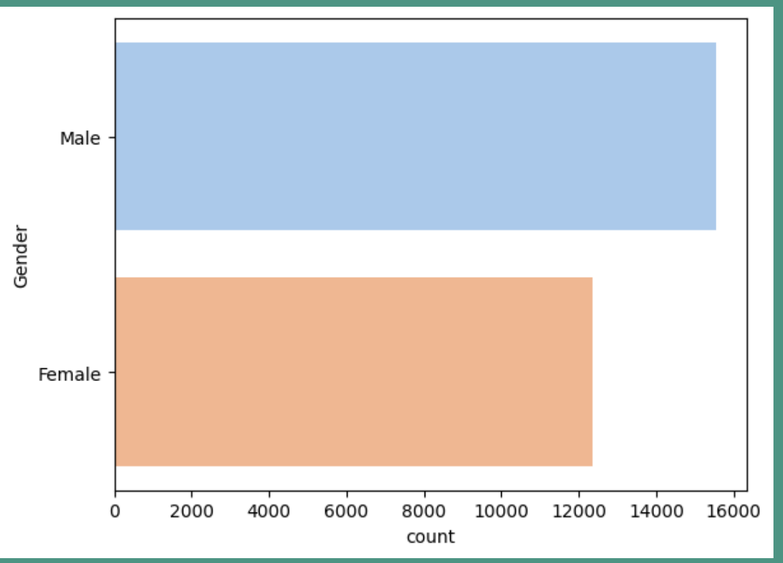
Work distribution chart

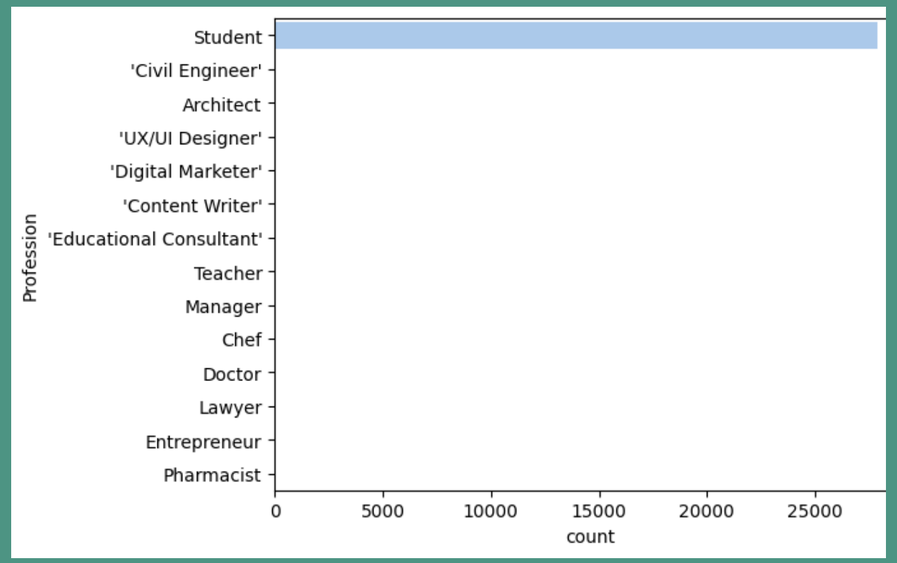
周鈺祥	陳胤宏
資料前處理	資料前處理
資料分布視覺化	建置模型
建置初步模型	性能評估
閱讀文獻方法	實驗結果視覺化
專案定義介紹等文件要求	資料集介紹
簡報設計排版製作	結論與模型限制

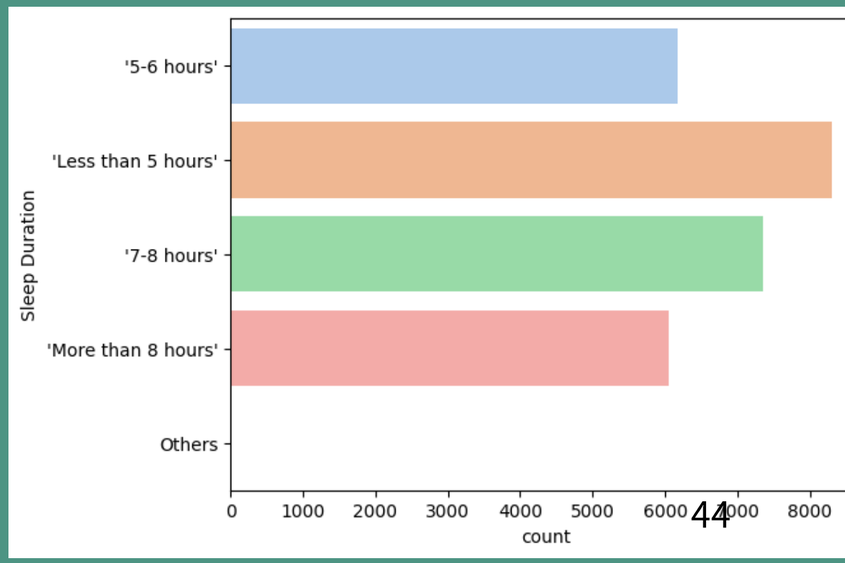
Appendix

1. Explore and visualize the data distribution
2. Clean the data by removing missing and erroneous values
3. Split the dataset into training (80%) and testing (20%) sets
4. Validate the performance of existing models (RF, SVM..)

Appendix







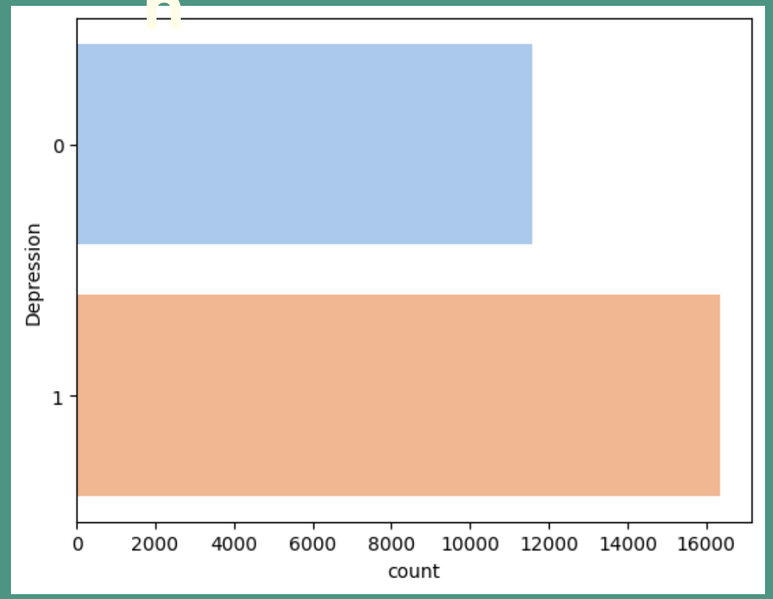
Gender

Profession

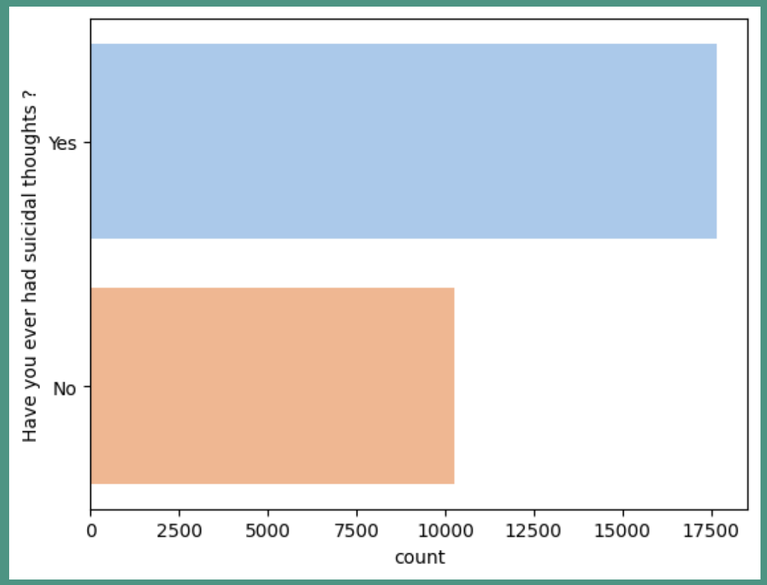
Sleep Duration

Appendix

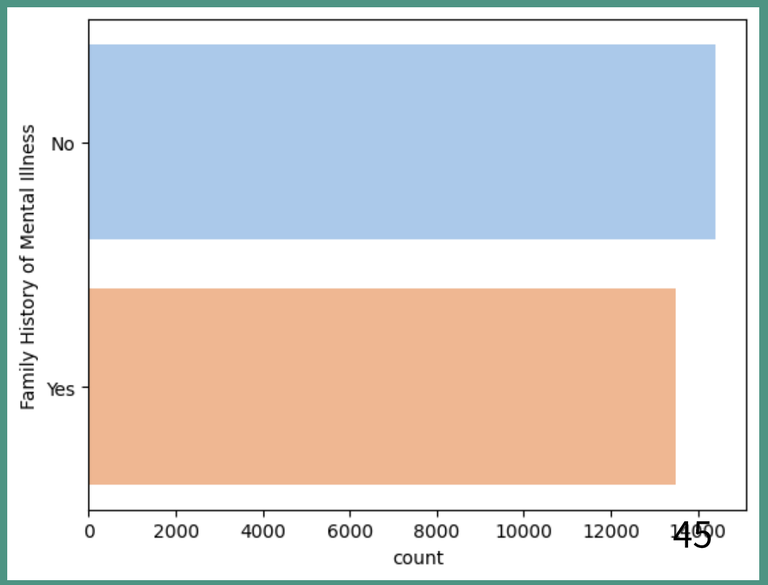
Depression



Have you ever had suicidal thoughts

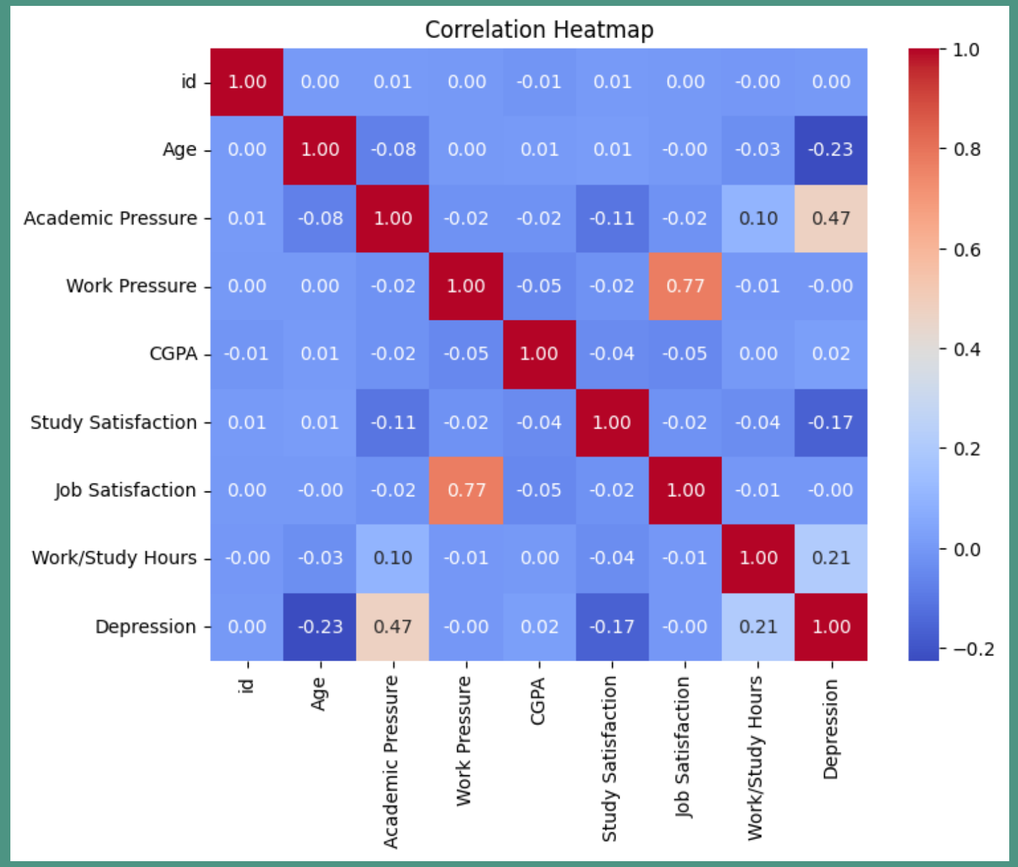


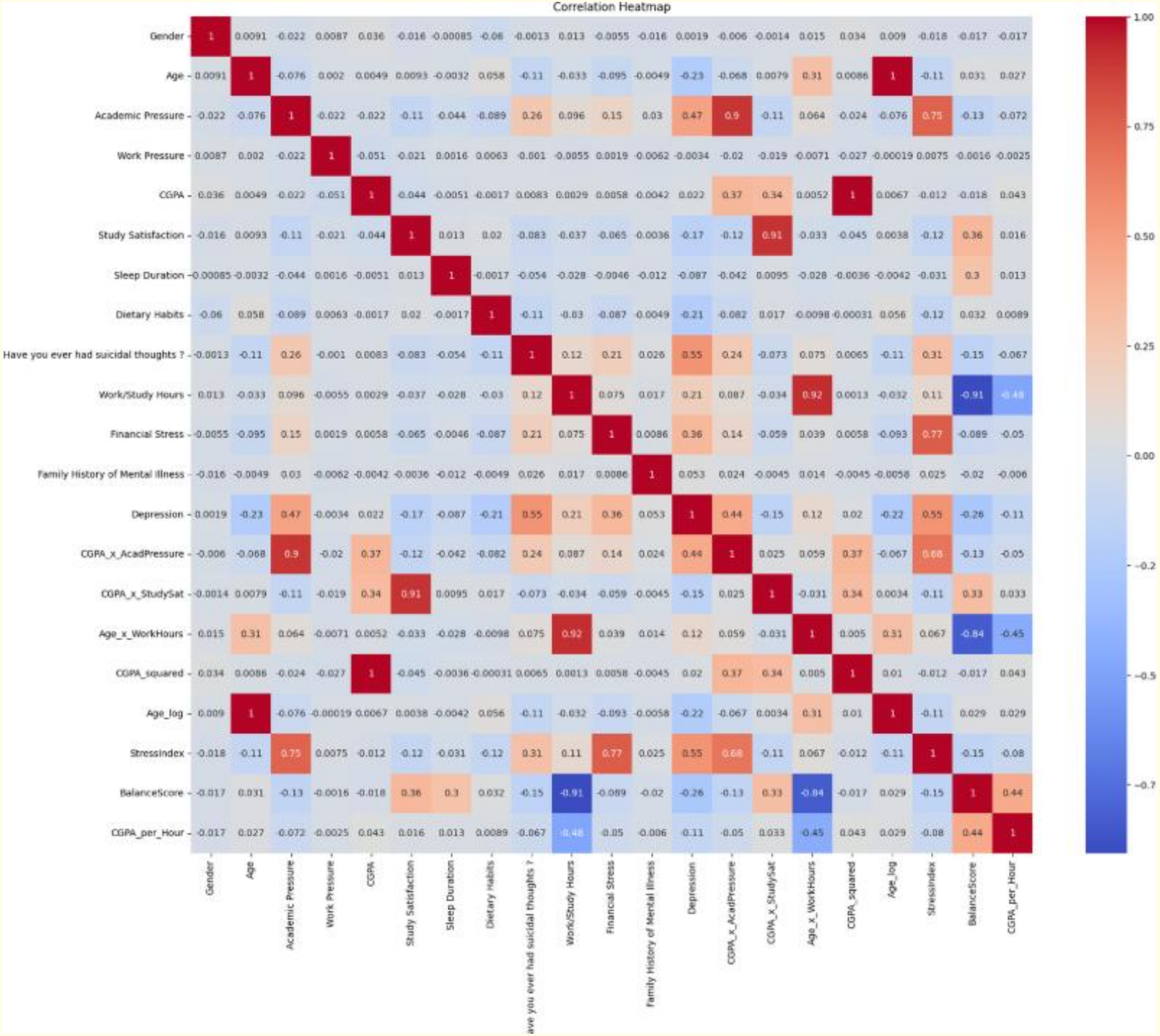
Family History of Mental Illness



Appendix

At this stage, only integer-type variables are included in the correlation calculation




Correlation Heatmap
Gender
Age
Academic Pressure
Work Pressure
CGPA
Study Satisfaction
Sleep Duration
Dietary Habits
Have you ever had suicidal thoughts ?
Work/Study Hours
Financial Stress
Family History of Mental Illness
Depression
CGPA_x_AcadPressure
CGPA_x_StudySat
Age_x_WorkHours
CGPA_squared
Age_log
StressIndex
BalanceScore
CGPA_per_Hour
1.00
0.75
0.50
0.25
0.00
−0.2
−0.5
−0.7

Student Depression Project

Gantt Chart

PROCESS	MARCH				APRIL				MAY	
	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2
Subject search										
Data understanding										
Data preprocess										
Select model										
Develope										
Evaluation and mantain										