DSP PROPOSAL

Student Depression

Analyzing Mental Health Trends & Predictors Among Students

Group A（Team 6）

312554010 周鈺祥　312554036 陳胤宏

March 26, 2025
10:00 a.m.

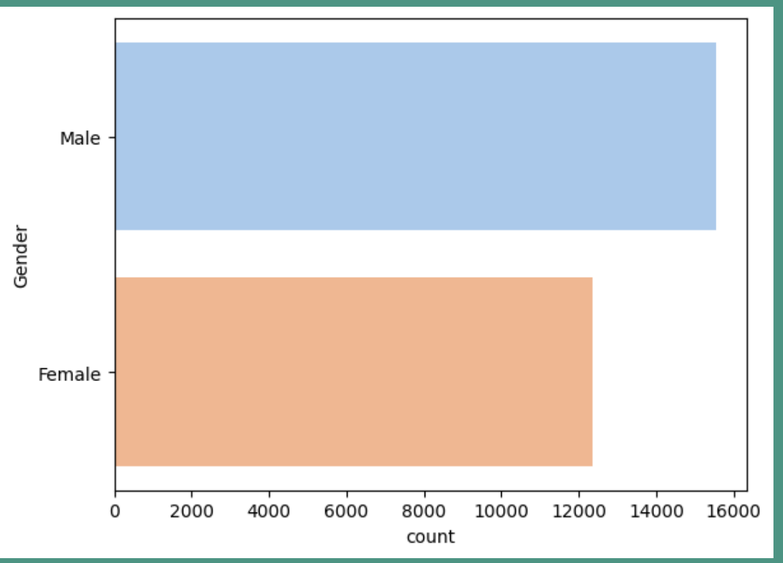
What we'll discuss later

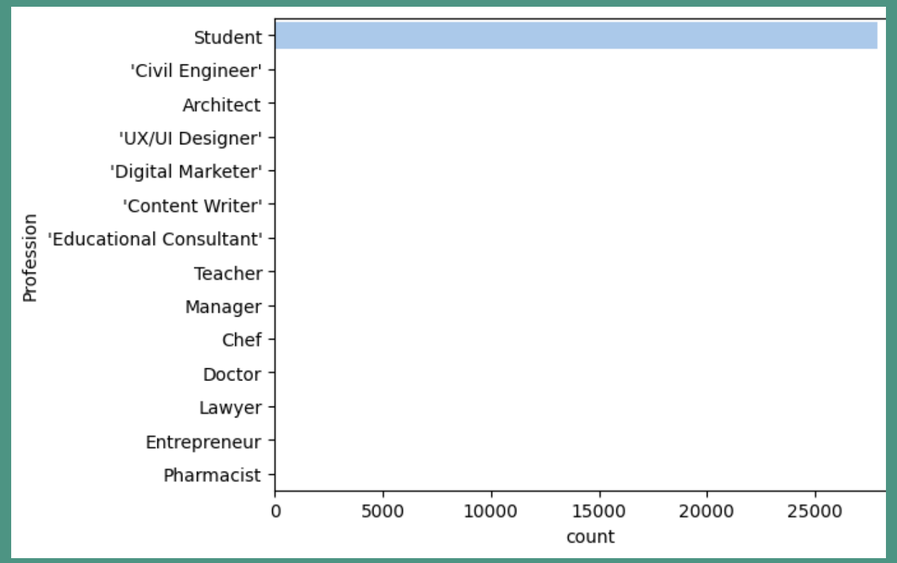
- → **Progress Update**
- → Introduction
- → Datasets
- → Problem Setting
- → Basic Work Plan
- → **Related Works**

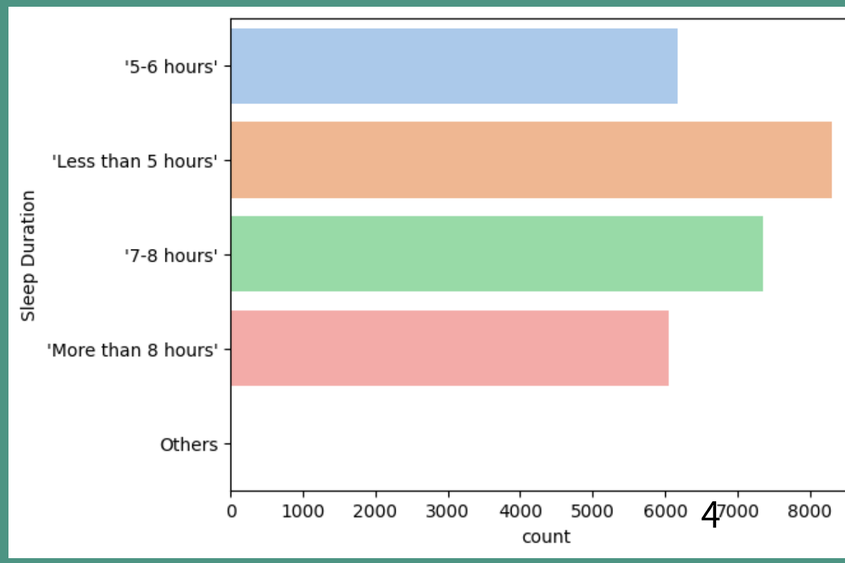
Progress Update

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..)

Progress Update



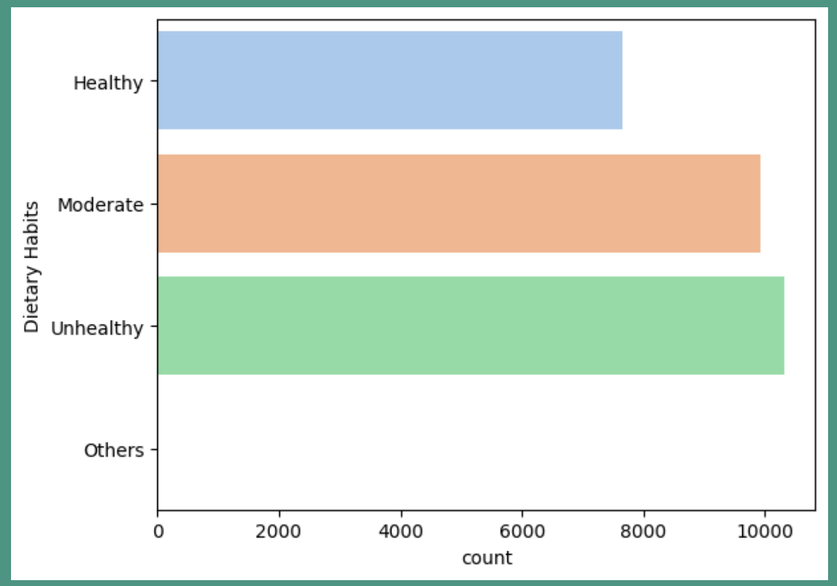




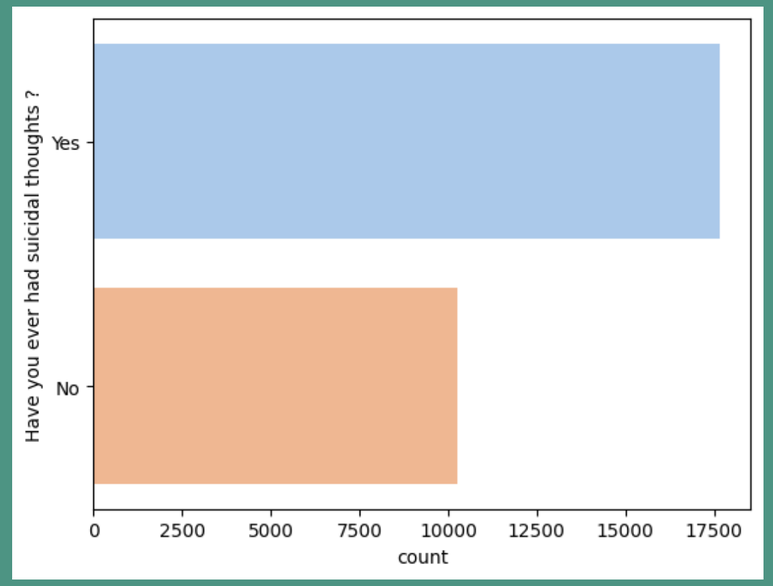
Gender **Profession** **Sleep Duration**

Progress Update

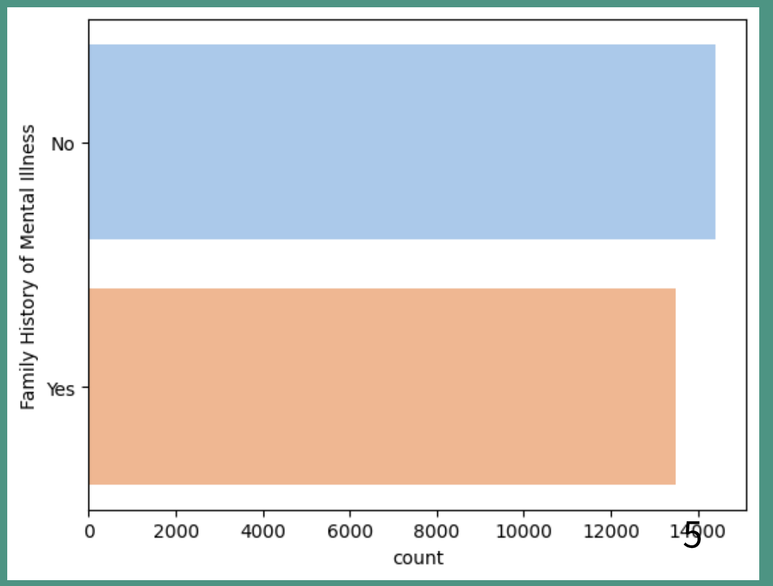
Dietary Habits



Have you ever had suicidal thoughts

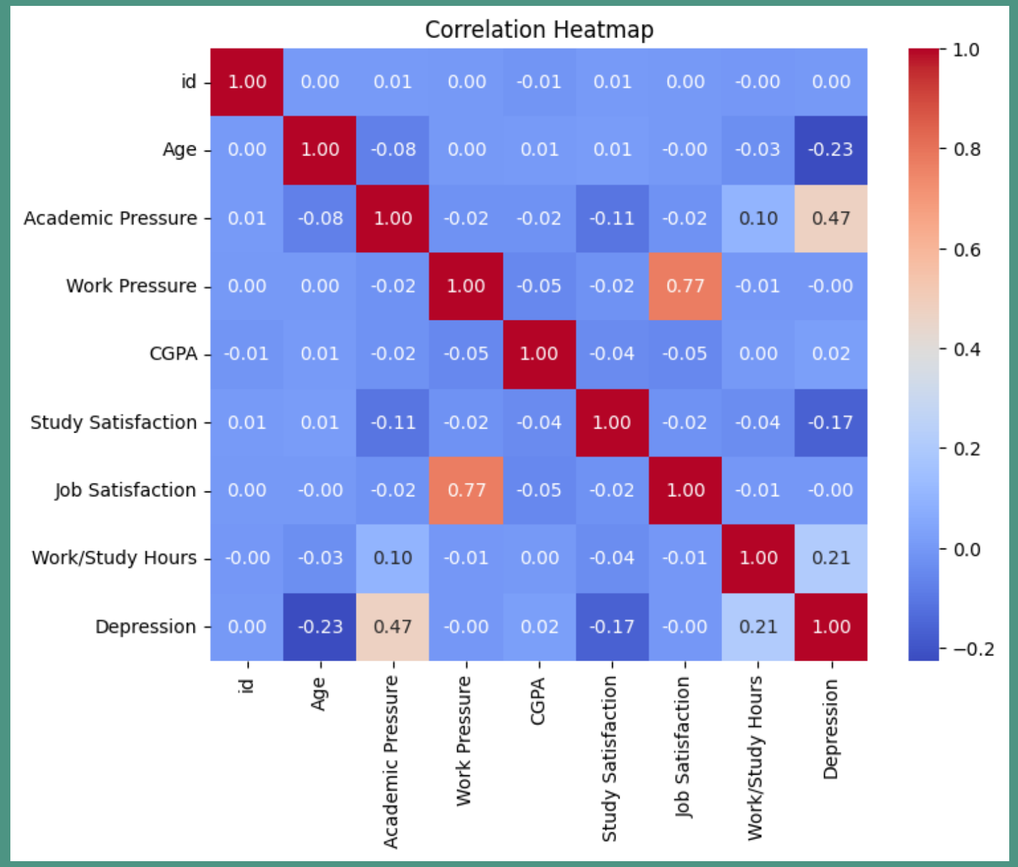


Family History of Mental Illness



Progress Update

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



Progress Update

RF	SVM	LR	DT
0.8439	0.8448	0.8470	0.7673

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

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

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

Expected 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.**

Data Description

Date Source

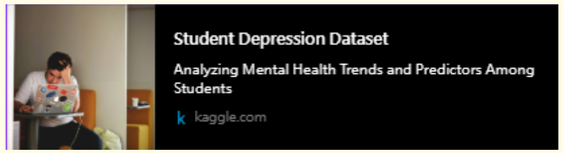
Kaggle

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

Problem Description

Train a transformer-based model using the dataset to predict whether an individual exhibits signs of depression

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.

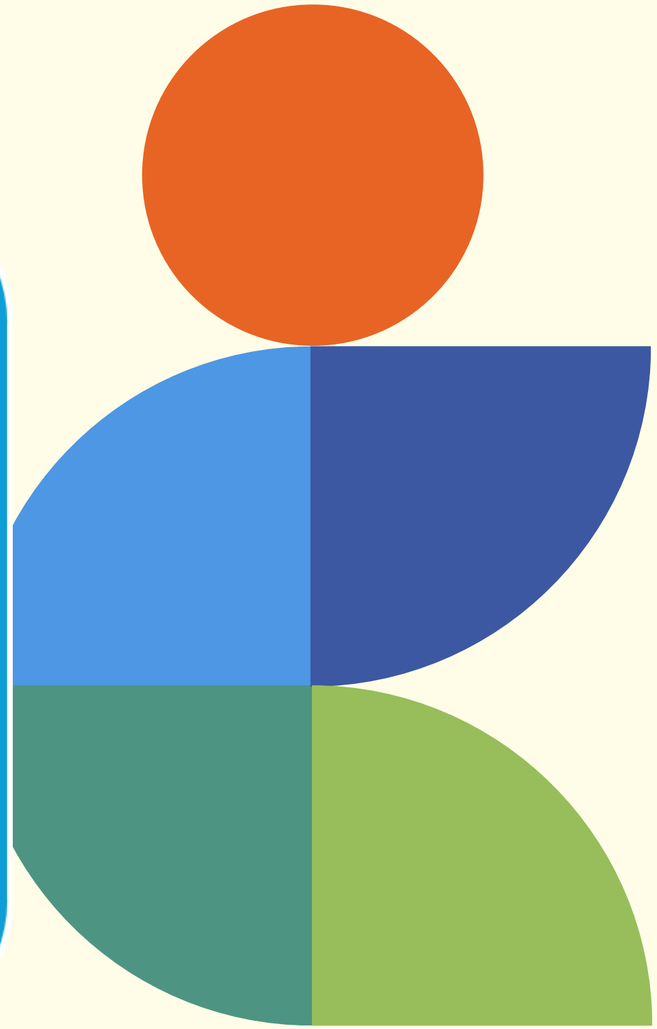
Evaluation Metrics

Accuracy

模型正確的分類率

P:Prediction
G:Ground truth
N:Total number of data

$$Accuracy = 1 - \frac{\sum_1^N (Pi - Gi)^2}{N}$$



Target Performance

Current rank

Random Forest Accuracy: 85.46%

Logistic Regression Accuracy: 85.53%

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

Target Performance

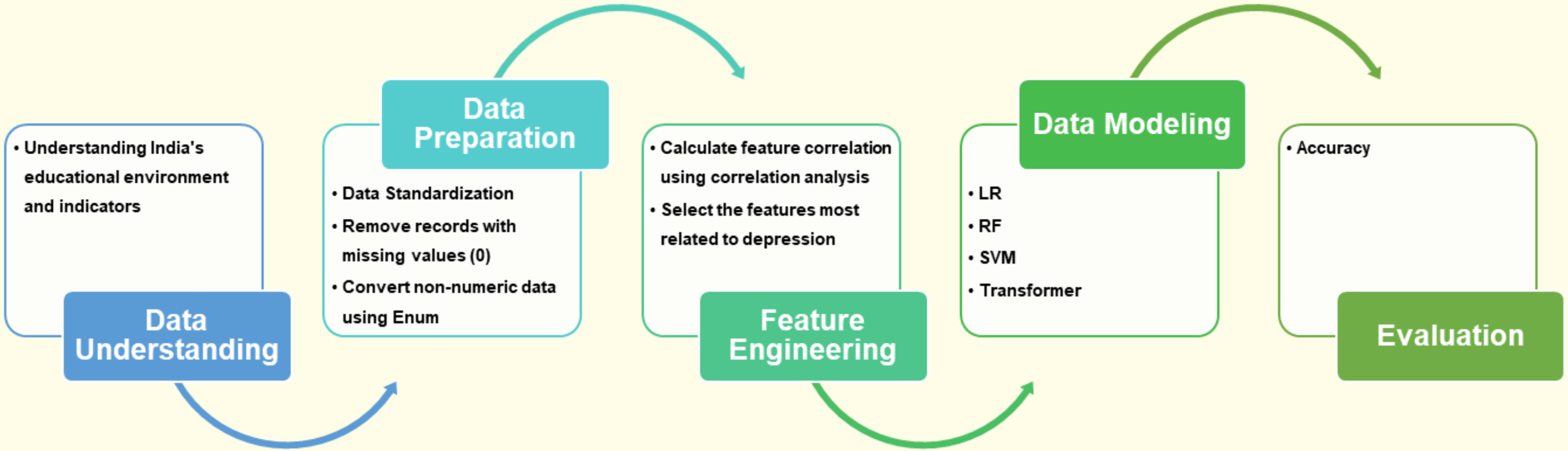
Reach 88% for Accuracy

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.

Basic Work Plan

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

Analysis workflow



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										

Related Works

Title	**Source**	**Year**	**Author**
Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis	JMIR Formative Research(https://pubmed.ncbi.nlm.nih.gov/35665695/)	2022	Radwan Qasrawi、Stephanny Paola Vicuna Polo、Diala Abu Al-Halawa、Sameh Hallaq , Ziad Abdeen
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
Identifying Adolescent Depression and Anxiety Through Real-World Data and Social Determinants of Health: Machine Learning Model Development and Validation	*JMIR Mental Health*, 2025; 12:e66665 (https://doi.org/10.2196/66665)	2025	Mamoun T. Mardini, Georges E. Khalil, Chen Bai, Aparna Menon DivaKaran, Jessica M. Ray
A multi-step water quality prediction model based on the Savitzky-Golay filter and Transformer optimized network	Springer Nature (https://link.springer.com/article/10.1007/s11356-023-29920-9)	2023	Ruiqi Wang、Ying Qi、 Qiang Zhang、Fei Wen
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

Summary

Title	**Method**	**Performance comparisons**	**Insights/Pitfalls**
Assessment and Prediction of Depression …	SVM, RF, DT, NB、ANN	Acc SVM:92.6，RF:92.4，NN:91.9，DT:88.5，NB:87.1	Insights:Experiences of violence, family support, academic performance, and PTSD are the most influential predictive factors. Pitfalls:Limited number of variables.
Prediction of adolescent depression …	DT, NB, SVM, AdaBoost, RF, MLP,RNN、LSTM	Overall Acc:0.64	Insights:Recursive Feature Elimination (RFE) effectively reduces feature dimensionality. Pitfalls:Severe issues with data imbalance and missing values.
Identifying Adolescent Depression …	*XGBoost*	AUC:0.81	Insights:Separate models for anxiety or depression perform better than combined models, likely due to reduced feature overlap and clearer symptom patterns. Pitfalls:Single geographic and demographic scope (Florida, USA) .Model generalizability to other populations and healthcare systems is uncertain.
A multi-step water quality prediction …	SG filter + Transformer	MAE:(T+1) 0.103	Insights:SG fiter is an annotative method to smooth the data. Pitfalls:Over-smoothing by the SG filter may suppress important short-term anomalies.
Identification of depressive symptoms …	SVM, MLP, RF, SLR	SVM/ P:0.920， SLR/ AUC:0.879，R:0.851，F1:0.868	Insights:Combining childhood and adolescent features significantly improves prediction. Pitfalls:Small sample size and limited generalizability.

Paper 1

D. Salas-Rueda, "Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis," JMIR Formative Research, vol. 6, no. 8, p. e37618, Aug. 2022

Input	Child Data (Total of 3984 individuals, aged 10-15 years, grades 5-9) Survey Questionnaire Data, including: • Socio-demographic variables (gender, age, parental education level, family income, etc.) • Mental health scales (DSRS for depression, GAD-7 for anxiety, PTSD scale) • Health behaviors (diet, exercise, sleep) • Social support (family, peers, school) • Experiences of family and school violence, academic performance, etc.
Process	Data Processing: • Handling missing values, standardization, and clustering Model Training and Validation: • Utilized five machine learning models (SVM, RF, ANN, DT, NB) • Performed cross-validation and parameter tuning (grid search) Feature Selection: • Used Random Forest to rank feature importance
Output	Classify schoolchildren as either "abnormally depressed/anxious" or not

Experiments

Data Source:
- **Academic year: 2013–2014**
- **Region: Public and refugee schools in the West Bank and East Jerusalem, Palestine**
- **Sample: 3,984 schoolchildren, aged 10–15, enrolled in grades 5 to 9**

Feature Variables:
- **Sociodemographic data (e.g., gender, grade level, family economic status)**
- **Health behaviors (e.g., diet, physical activity)**
- **Social support**
- **Psychological conditions (e.g., PTSD)**
- **Experiences of family and school violence**

Target Variables:
- **Depression (normal / abnormal)**
- **Anxiety (normal / abnormal)**

Splitting:
- **Training: 70%**
- **Testing: 20%**
- **Validation: 10%**

Model Training Parameters:
- **SVM : RBF kernel · C=20 · γ=0.001**
- **- RF : 1000 trees · max_depth=5**
- **- NN : 500 neurons · logistic activation · max_iter=500**

Algorithm:
- **Support Vector Machine (SVM)**
- **Random Forest (RF)**
- **Neural Network (ANN)**
- **Decision Tree (DT)**
- **Naive Bayes (NB)**

Model, mental Health condition	AUC[a], %	CA[b], %	Error rate, %	F1-score[c], %	Precision, %	Recall, %
Decision tree						
Depression	86.7	88.5	88.5	88.5	88.5	86.7
Anxiety	73.7	74.4	74.1	74.2	74.4	73.7
Support vector machine						
Depression	96.8	92.5	92.6	93.7	92.5	96.8
Anxiety	82.1	76.4	76.5	76.8	76.4	82.1
Random forest						
Depression	97.2	92.4	92.4	93.3	92.4	97.2
Anxiety	86.8	78.6	78.4	78.5	78.6	86.8
Artificial neural network						
Depression	96.8	91.9	91.9	92.3	91.9	96.8
Anxiety	84	75.9	75.7	75.7	75.9	84
Naive Bayes						
Depression	95.5	86.9	87.1	89.9	86.9	95.5
Anxiety	82.3	73	72.7	72.8	73	82.3

Paper 2

A. Yoo et al., "Prediction of Adolescent Depression from Prenatal and Childhood Data from ALSPAC Using Machine Learning," Scientific Reports, vol. 14, no. 23282, Oct. 2024, doi: 10.1038/s41598-024-72158-9.

Input	Data Source: ALSPAC (Avon Longitudinal Study of Parents and Children) Sample Size: 8467 children and their parents Feature Types: 885 features from pregnancy to the child's age of 10, including: • Biological characteristics (BMI, physiological indicators, maternal health during pregnancy) • Socioeconomic factors (family income, parental education, neighborhood deprivation) • Psychological factors (maternal and child depression levels, self-esteem, self-control) • Educational and cognitive development (academic performance, numeracy skills, school evaluation) • Emotional and interpersonal relationships (peer relationships, parent-child relationships, life events)
Process	• Data Cleaning and Transformation:Handling missing values (KNN, MICE, MissForest) • Feature scaling and categorical encoding • Oversampling to address class imbalance (SMOTE) • Feature Selection:Pearson correlation coefficient • Recursive Feature Elimination (RFE)
Output	Predicting whether an individual will experience depression at any time point between the ages of 12 and 18

Experiments

Data Source:

- **ALSPAC (Avon Longitudinal Study of Parents and Children), a large UK-based birth cohort study**
- **8,467 participants**
- **Feature data collected from prenatal to age 10**
- **Goal: Predict whether adolescents develop depression between ages 12–18**

Feature Variables:

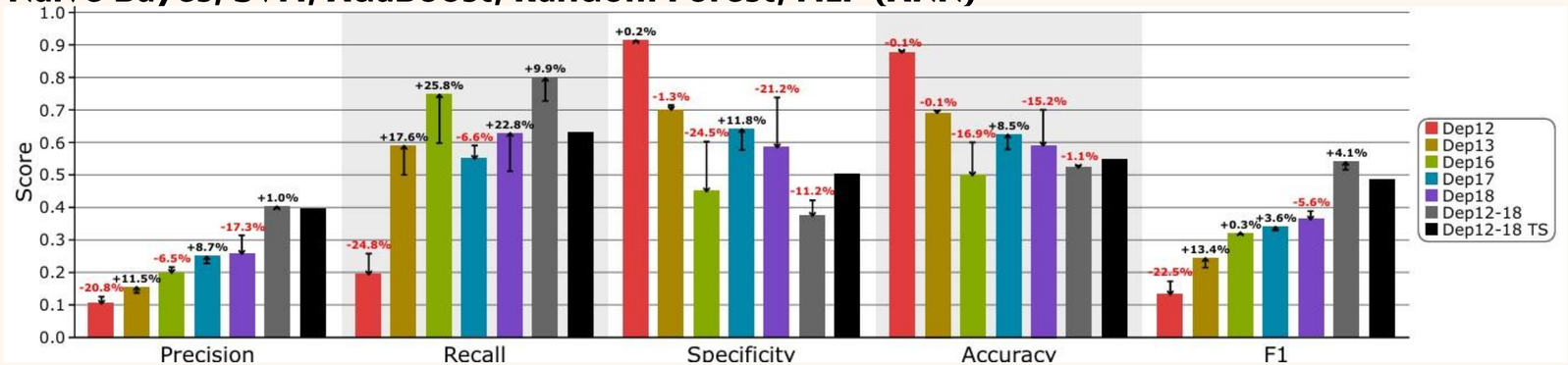
- **Original dataset contained 6,163 features**
- **After filtering, 885 features remained**
- **Final valid features for time series format: 377**

Target Variable:

- **Predict whether adolescents exhibited depression at ages 12, 13, 16, 17, or 18, or any point between**
- **Depression measured using the SMFQ (Short Mood and Feelings Questionnaire)**
- **Score ≥ 12 was classified as "depressed"**

Algorithm:

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



Paper 3

J. V. Osborne, H. B. Rollins, B. L. Kronick, N. R. Valdez, and L. S. Mertz, "Identifying adolescent depression and anxiety through real-world data and social determinants of health: Machine learning model development and validation," JMIR Ment. Health, vol. 12, no. 1, p. e66665, Jan. 2025. [Online]. Available: L. S. Mertz, H. B. Rollins, B. L. Kronick, J. V. Osborne, and N. R. Valdez, "Identifying adolescent depression and anxiety through real-world data and social determinants of health: Machine learning model development and validation," JAMA Network Open, vol. 6, no. 6, pp. e2319391, 2023.

Input	• Electronic health record (EHR) data (diagnoses, medications, procedures, laboratory results) • Regional social determinants of health (SDoH) data (e.g., income, education, internet access) • Geographic information (GEOID)
Process	Feature selection methods: Chi-squared test, correlation analysis, LASSO
Output	Three prediction models: anxiety only, depression only, and comorbid anxiety and depression

Experiments

Source: 52,054 adolescent electronic health records (ages 10–17) from the University of Florida Integrated Data Repository (UF-IDR)

- **Case Counts:**
- **12,572 with anxiety**
- **7,812 with depression**
- **14,019 with comorbid anxiety and depression**

Feature Types

- **Individual-level features: age, sex, race, diagnoses, medications, vital signs, medical procedures**
- **Social-level features: block-level social determinants of health (SDoH) provided by the National Historical Geographic Information System (e.g., education level, poverty rate)**
- **Special indicators: ADI (Area Deprivation Index)**

Preprocessing

- **Missing values imputed using K-nearest neighbors (KNN)**
- **Numerical features standardized**
- **Categorical features encoded**
- **Outliers replaced with median values**

模型預測目標	含 SDoH	僅 ADI	無 SDoH
焦慮	0.80	0.79	0.78
憂鬱	0.81	0.81	0.81
焦慮或憂鬱	0.78	0.78	0.78

Model

- **Built 9 XGBoost models based on:**
- **3 prediction tasks: anxiety only, depression only, comorbid cases**
- **3 feature settings: full SDoH, ADI only, no SDoH**
- **Evaluation: 5x5 nested cross-validation with grid search for hyperparameter tuning**

Paper 4

R. Wang, Y. Qi, Q. Zhang, and F. Wen, "A multi-step water quality prediction model based on the Savitzky-Golay filter and Transformer optimized network," Environmental Science and Pollution Research, published online: Sep. 28, 2023. [Online]. Available: https://doi.org/10.1007/s11356-023-30041-z

Input	Original water quality time series data (from 4 monitoring points in the Lanzhou section of the Yellow River, China, 2018–2019, hourly dissolved oxygen data, a total of 39,026 data points)
Process	Data smoothing and denoising using the Savitzky-Golay (SG) filter Min-max normalization processing Slicing using a sliding time window
Output	Multi-step water quality prediction results with higher accuracy than other benchmark models (capable of predicting 3 steps or more). The results can serve as a basis for watershed pollution management and water quality control

Experiments

Data Processing:
Missing values imputed using KNN Outliers detected and replaced using Isolation Forest Min-max normalization applied Dataset

Splitting:
Training set: 80% Validation set: 10% Testing set: 10%

Model Training Parameters:
Time step (input sequence length): 24 hours Batch size: 200 Learning rate: 0.05, learning rate decays by a factor of 0.7 every 20 epochs Number of epochs: 60 Adam optimizer used, Dropout for overfitting prevention Number of multi-head attention heads: 5

Table 3 Performance comparison of each water quality prediction model at different time steps

Models	MAE			RMSE			NSE		
	T+1	T+2	T+3	T+1	T+2	T+3	T+1	T+2	T+3
LSTM	0.178	0.227	0.267	0.142	0.185	0.232	0.873	0.855	0.812
Seq2Seq-LSTM	0.163	0.166	0.234	0.144	0.161	0.213	0.892	0.879	0.851
Transformer(MSE)	0.135	0.151	0.173	0.123	0.130	0.161	0.914	0.874	0.853
Transformer(DILATE)	0.121	0.142	0.162	0.114	0.125	0.149	0.928	0.893	0.870
SG-Transformer(MSE)	0.118	0.134	0.171	0.113	0.134	0.137	0.924	0.906	0.877
SG-Transformer	0.103	0.101	0.153	0.076	0.097	0.123	0.947	0.921	0.891

Paper 5

X. Liu, R. Cang, Z. Zhang, P. Li, H. Wu, W. Liu, and S. Li, "Identification of depressive symptoms in adolescents using machine learning combining childhood and adolescence features," BMC Public Health, vol. 25, no. 264, 2025. [Online]. Available: https://doi.org/10.1186/s12889-025-21506-z

Input	Source data: China Family Panel Studies (CFPS) Participants: Adolescents aged 13–15, traced back to age 6–9 (2012 or 2014 data) Final sample size: 505 adolescents Features: 39 variables, including: • Childhood features (9 total): family relationships, personality, parental behavior • Adolescent features (30 total): behavior/lifestyle, psychological perception, family interaction, academic pressure, parenting • Demographic features: gender, age, residence, health self-assessment, number of family members
Process	Preprocessing: • Data split into training set (n = 403) and test set (n = 102) • Borderline-SMOTE used to handle class imbalance • All features normalized using Z-score Feature Selection: • mRMR (Maximum Relevance Minimum Redundancy) used for ranking and initial filtering (top 20 features) • Features added one-by-one into four ML models with 5-fold cross-validation for optimal subset selection
Output	Predicting whether an adolescent has depressive symptoms

Experiments

Dataset: China Family Panel Studies (CFPS) from 2020 (adolescents) and 2012/2014 (childhood)

Sample: 505 adolescents (excluding missing values and those with diagnosed depression)

Split: Random 80:20 for training vs test

Class Imbalance: Borderline-SMOTE

Normalization: Z-score

Feature Selection: mRMR + cross-validation

Algorithms: SVM, MLP, RF, SLR

Hyperparameter Tuning:

- **RF: Random Search (estimators, depth, split size, leaf size)**
- **MLP/SVM: GridSearchCV (layers, activation, C, γ, etc.)**

Table 2 Performance of all models predicting depressive symptoms on the training and test sets

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

Conclusion

Referenced Methods

1. SMOTE / Borderline SMOTE
2. Recursive Feature Elimination (RFE) effectively reduces feature dimensionality
3. Performed cross-validation and parameter tuning (grid search)
4. Transformer model method

Conclusion

While the overall number of features is relatively small, reducing the input to a selected subset may limit the model's learning capacity and lead to underfitting. Additionally, it may cause the model to overly depend on specific features, making it sensitive to minor anomalies and potentially impairing prediction accuracy.

Thank you!

Student Depression
DSP Proposal

Group A

312554010 周鈺祥 312554036 陳胤宏

March 26, 2025
10:00 a.m.