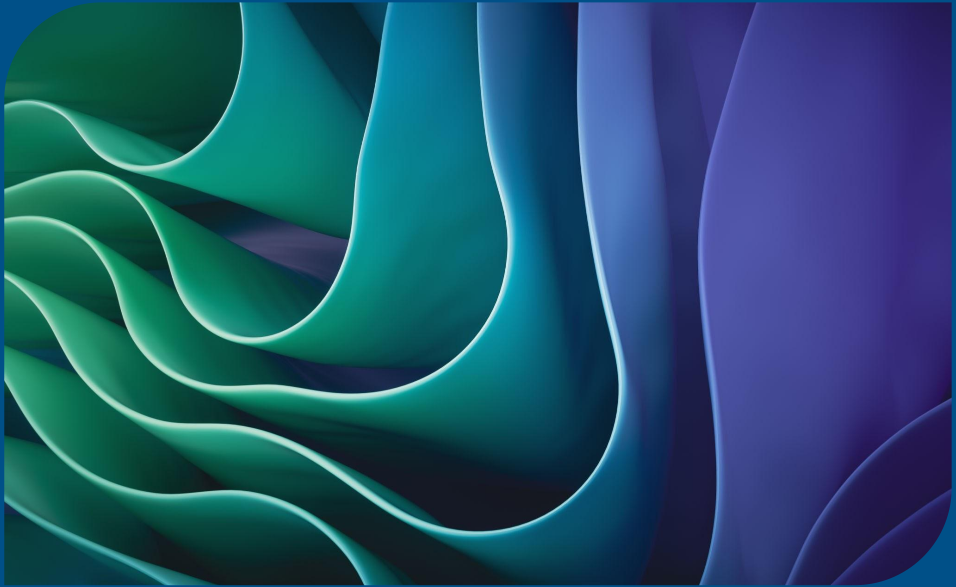
12.05.24

Exploring Mental Health Data

Viktoriia Kachanovska



1. **Data Overview**
2. **Data Cleaning**
3. **EDA**
4. **Model**
5. **Results, Analysis and Conclusions**

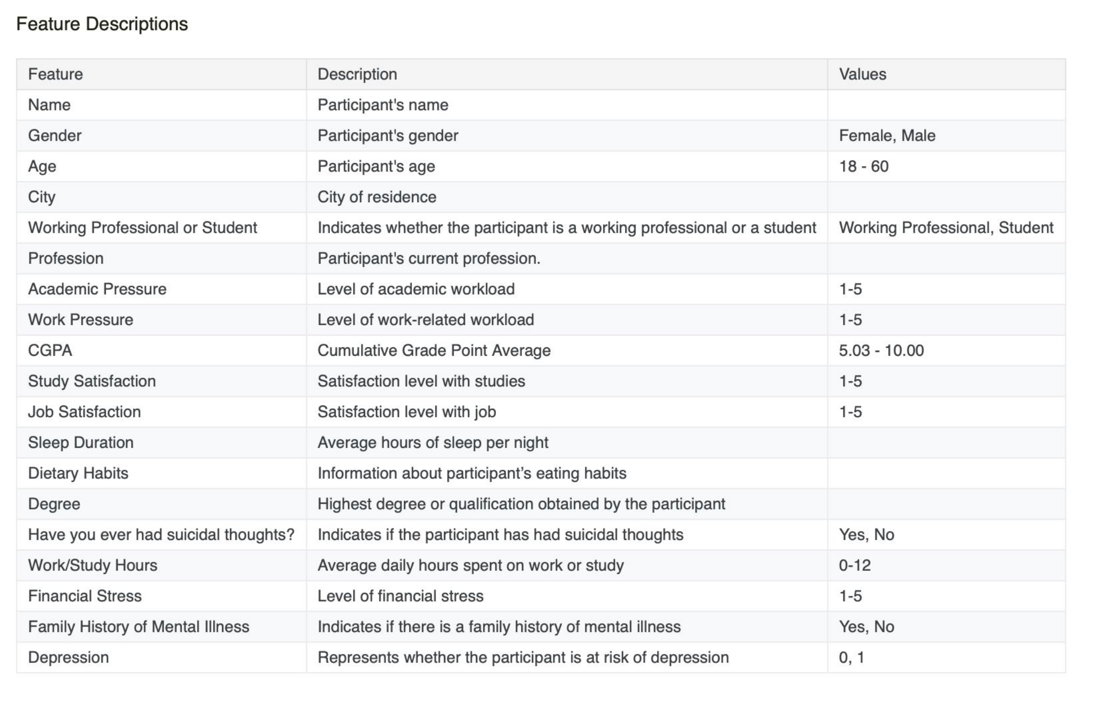
The goal of this project is to use data from a mental health survey to explore factors that may cause individuals to experience depression. This project uses supervised machine learning to address a binary classification task, where the target variable is Depression.

The dataset explores various factors such as sleep duration, dietary habits, academic pressure, professional satisfaction, etc. to identify contributors to depression.

The dataset for this project (both train and test) was generated from a deep learning model trained on the Depression Survey/Dataset for Analysis dataset.

train dataset rows: 140700, columns: 20

test dataset rows: 93800, columns: 19



Feature Descriptions

Feature	Description	Values
Name	Participant's name	
Gender	Participant's gender	Female, Male
Age	Participant's age	18 - 60
City	City of residence	
Working Professional or Student	Indicates whether the participant is a working professional or a student	Working Professional, Student
Profession	Participant's current profession.	
Academic Pressure	Level of academic workload	1-5
Work Pressure	Level of work-related workload	1-5
CGPA	Cumulative Grade Point Average	5.03 - 10.00
Study Satisfaction	Satisfaction level with studies	1-5
Job Satisfaction	Satisfaction level with job	1-5
Sleep Duration	Average hours of sleep per night	
Dietary Habits	Information about participant’s eating habits	
Degree	Highest degree or qualification obtained by the participant	
Have you ever had suicidal thoughts?	Indicates if the participant has had suicidal thoughts	Yes, No
Work/Study Hours	Average daily hours spent on work or study	0-12
Financial Stress	Level of financial stress	1-5
Family History of Mental Illness	Indicates if there is a family history of mental illness	Yes, No
Depression	Represents whether the participant is at risk of depression	0, 1

Age: Mean age is approximately 40 years with standard deviation of 12, which indicates the mix of younger and middle-aged individuals. Based on distribution there is large concentration of the youngest group (18-20 years).

Academic and Work Pressure: The mean value is 3.14 for Academic Pressure and 2.99 for Work Pressure. But approximately 80% of the dataset is missing for Academic Pressure because this feature applies only for students.

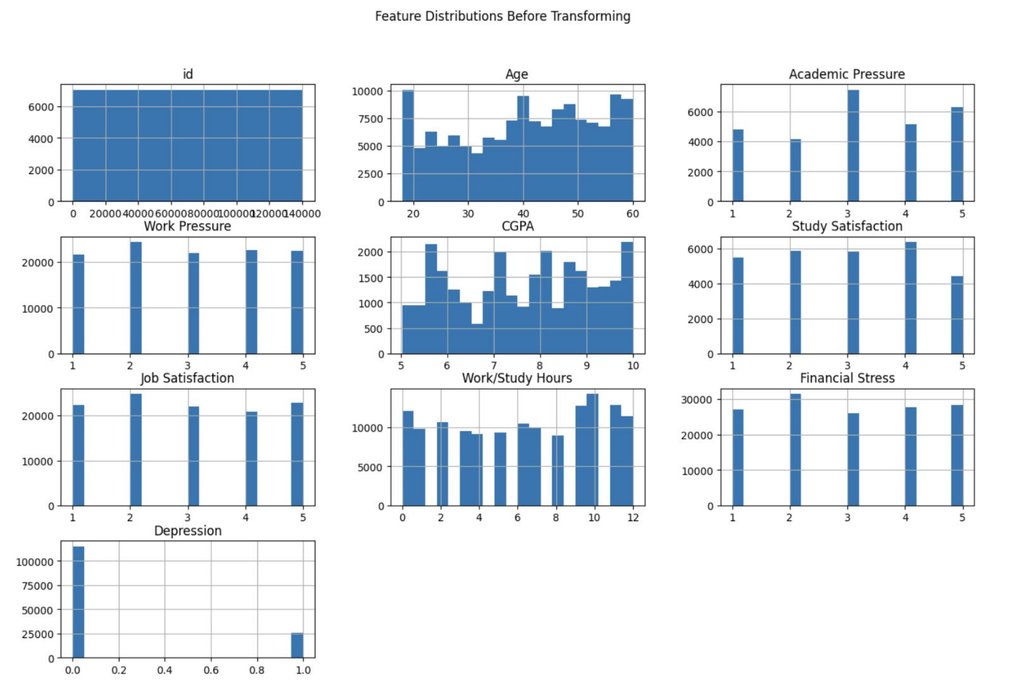
CGPA: The mean value is 7.66 (out of 10).

Study and Job Satisfaction: The mean value is near 3 on scale of 1 to 5. But approximately 80% of the dataset is missing for Study Satisfaction because this feature applies only for students.

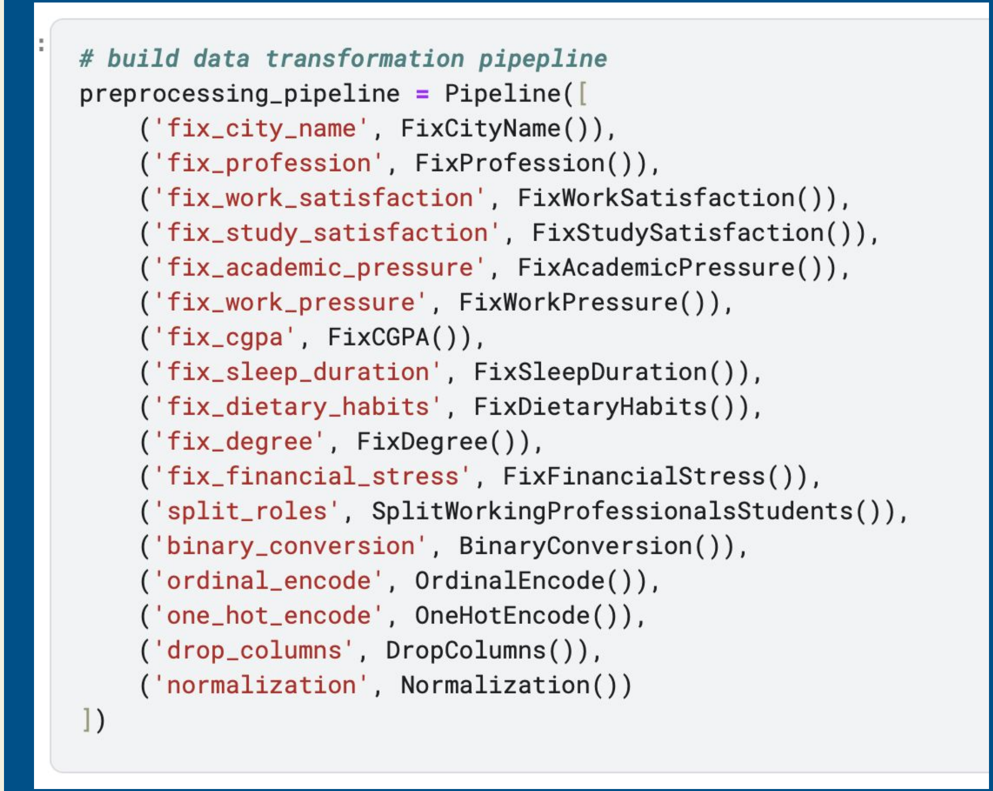
Work/Study Hours: The mean Work/Study Hours is 6.25, with a maximum of 12 hours and a median of 6 hours.

Financial Stress: Financial Stress ranges from 1 to 5 with a mean of 2.99. The even distribution (median 3.0) suggests financial pressure may be fairly balanced across the dataset.

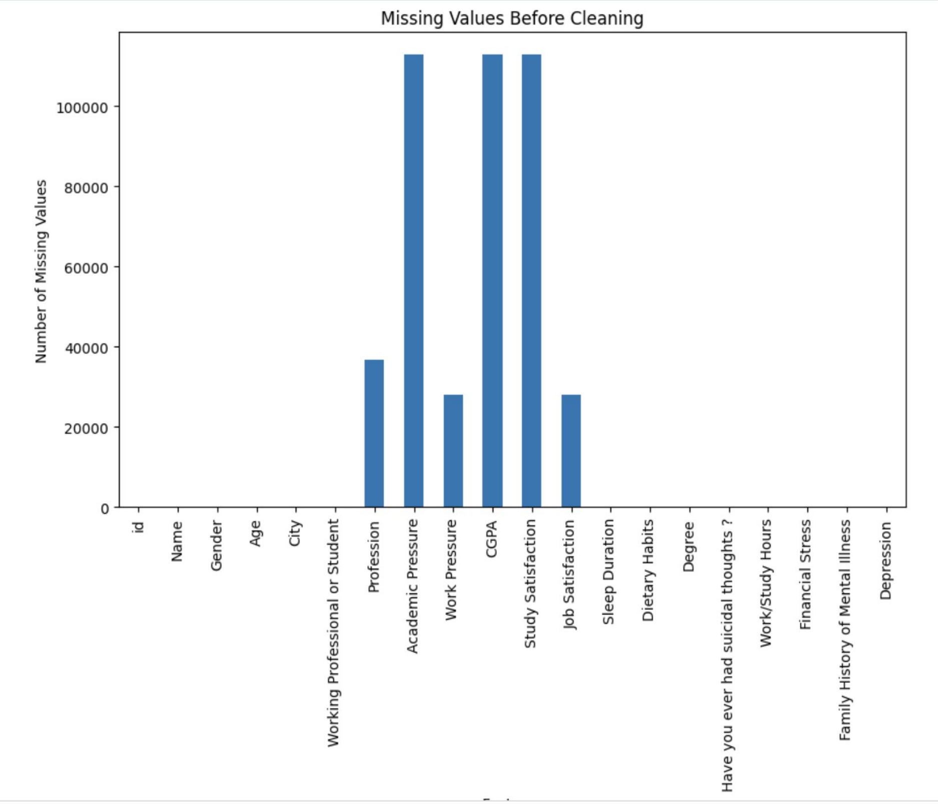
Depression: About 18% of participants have reported symptoms of depression. The distribution of the Depression is imbalanced, requiring balancing techniques like oversampling or class-weight adjustments for predictive modeling.

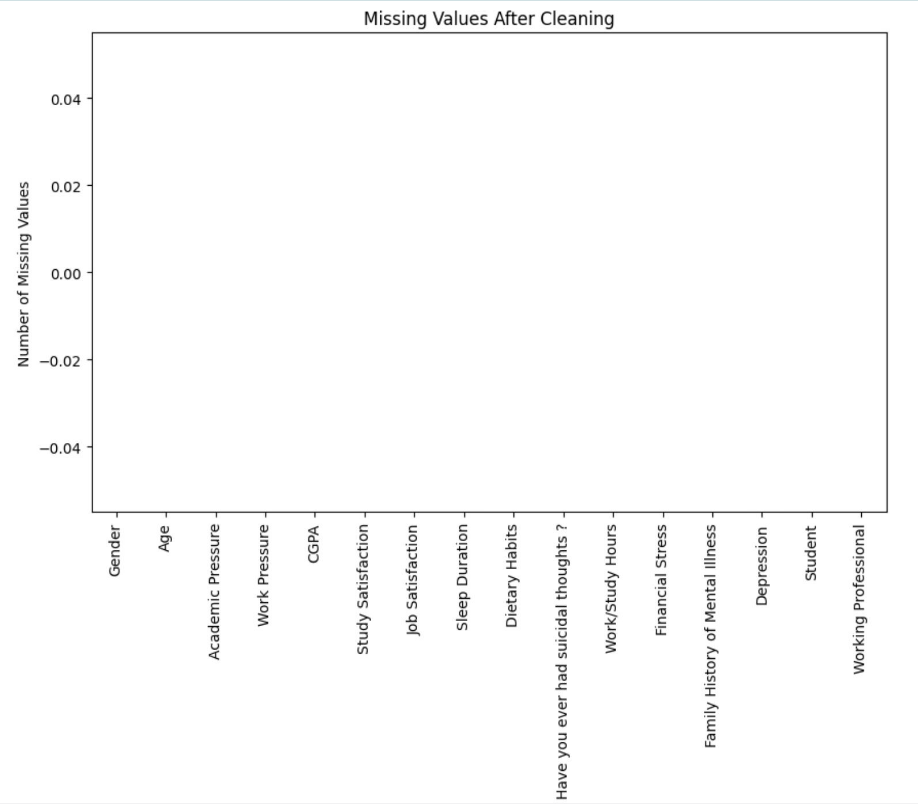


The data cleaning process includes standardization, imputing, and transformation raw input data for a mental health-related dataset. Were created custom transformers for handling missing values, standardization of categorical variables, normalization of numerical features, and feature engineering.



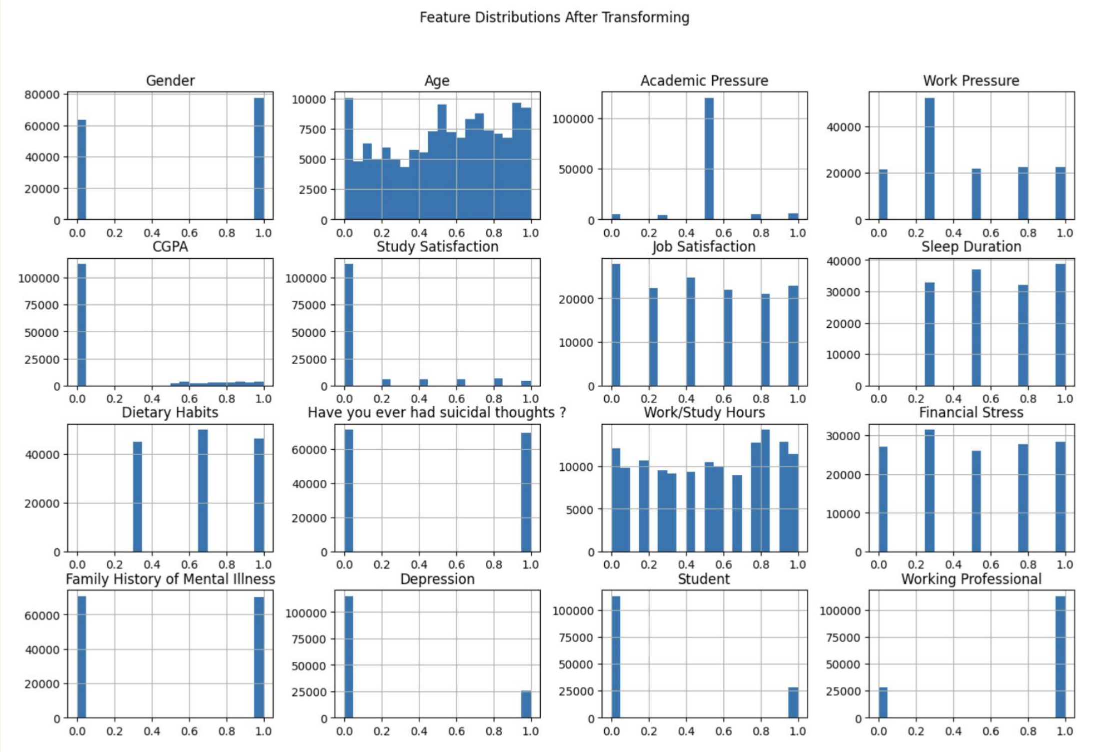
```
# build data transformation pipepline
preprocessing_pipeline = Pipeline([
    ('fix_city_name', FixCityName()),
    ('fix_profession', FixProfession()),
    ('fix_work_satisfaction', FixWorkSatisfaction()),
    ('fix_study_satisfaction', FixStudySatisfaction()),
    ('fix_academic_pressure', FixAcademicPressure()),
    ('fix_work_pressure', FixWorkPressure()),
    ('fix_cgpa', FixCGPA()),
    ('fix_sleep_duration', FixSleepDuration()),
    ('fix_dietary_habits', FixDietaryHabits()),
    ('fix_degree', FixDegree()),
    ('fix_financial_stress', FixFinancialStress()),
    ('split_roles', SplitWorkingProfessionalsStudents()),
    ('binary_conversion', BinaryConversion()),
    ('ordinal_encode', OrdinalEncode()),
    ('one_hot_encode', OneHotEncode()),
    ('drop_columns', DropColumns()),
    ('normalization', Normalization())
])
```


Missing Values Before Cleaning
Number of Missing Values
0
20000
40000
60000
80000
100000
id
Name
Gender
Age
City
Working Professional or Student
Profession
Academic Pressure
Work Pressure
CGPA
Study Satisfaction
Job Satisfaction
Sleep Duration
Dietary Habits
Degree
Have you ever had suicidal thoughts ?
Work/Study Hours
Financial Stress
Family History of Mental Illness
Depression

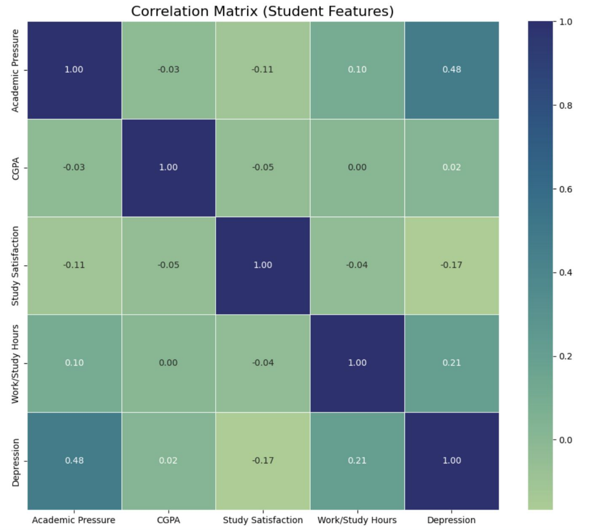

Missing Values After Cleaning
Number of Missing Values
0.04
0.02
0.00
−0.02
−0.04
Gender
Age
Academic Pressure
Work Pressure
CGPA
Study Satisfaction
Job Satisfaction
Sleep Duration
Dietary Habits
Have you ever had suicidal thoughts ?
Work/Study Hours
Financial Stress
Family History of Mental Illness
Depression
Student
Working Professional

Gender, Depression, Working Professional, Student, CGPA, Study Satisfaction has a significant imbalance, with one class being much more prevalent.

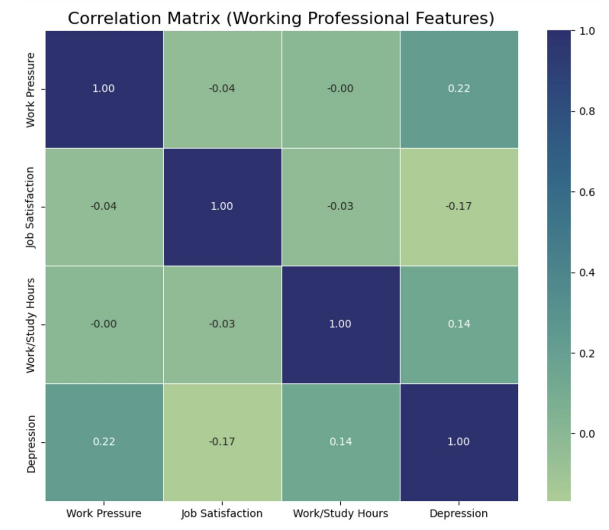
Age demonstrates a relatively uniform distribution across the range, indicating effective scaling.



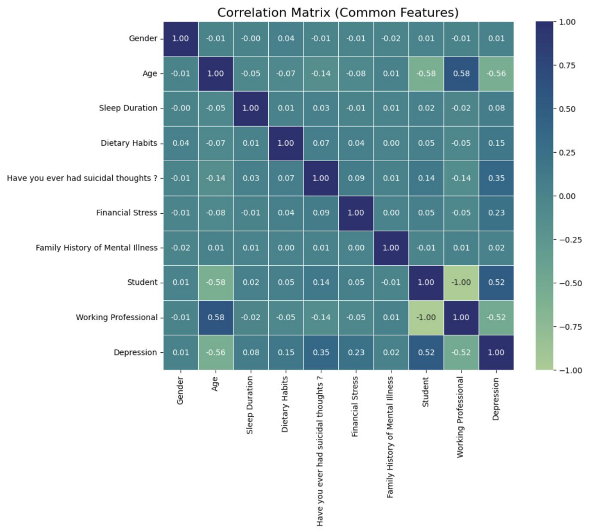
Correlation Matrices



Correlation Matrix (Student Features)



Correlation Matrix (Working Professional Features)



Correlation Matrix (Common Features)

Strong Negative Correlation:

Age and Depression (-0.56), suggesting that younger individuals might have higher depression scores.

Working Professional and Depression (-0.52), suggesting that working professional are less likely to experience depression.

Strong Positive Correlation:

Academic Pressure and Depression (0.48) shows that high academic pressure could be a factor contributing to depression.

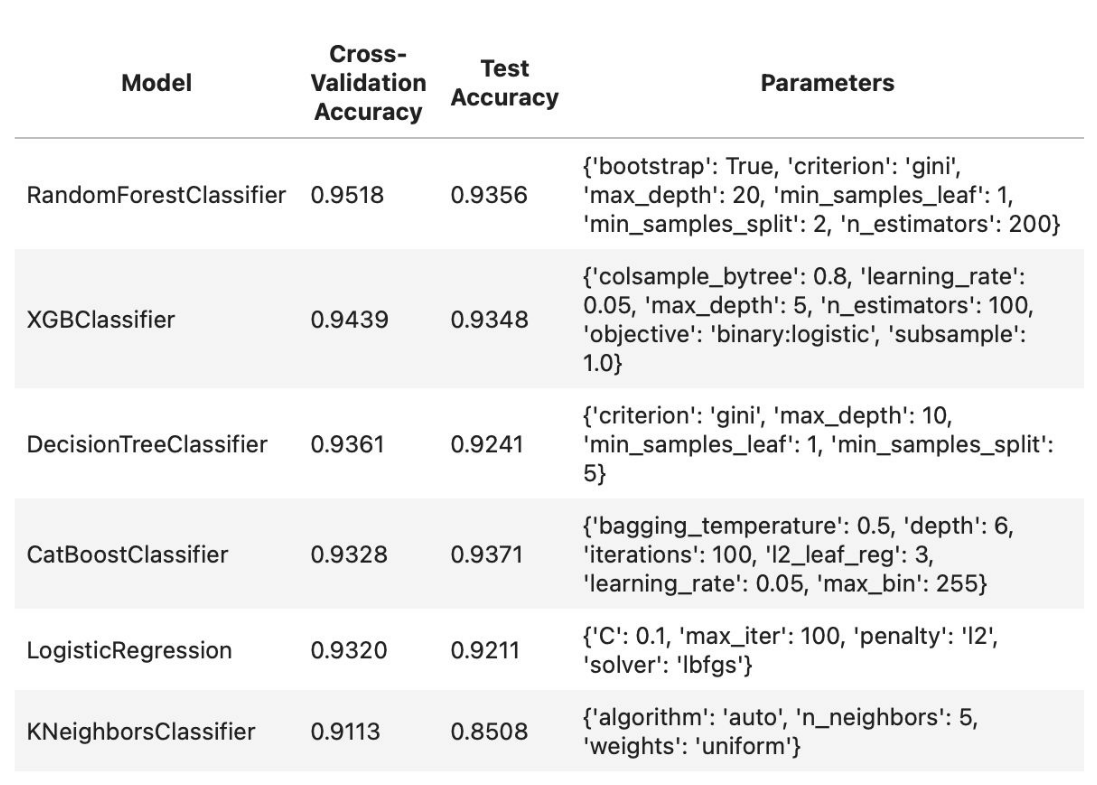
Student and Depression (0.52), suggesting that Students are more likely to experience depression.

From the table provided, was selected for **VotingClassifier** ensemble:

RandomForestClassifier: High cross-validation accuracy (0.9518) and test accuracy (0.9356).

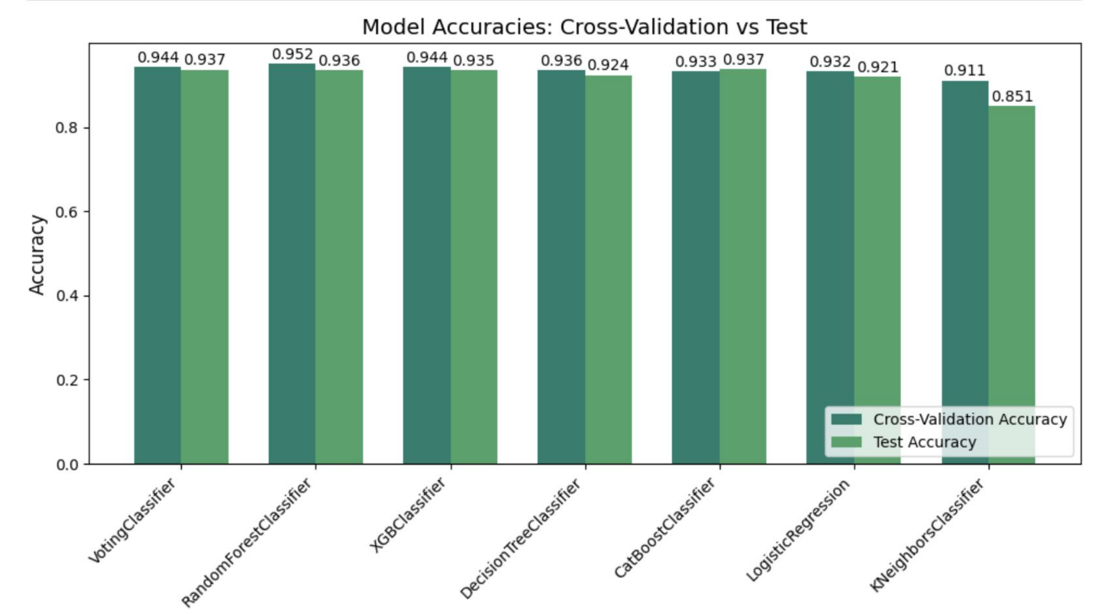
CatBoostClassifier: Slightly lower cross-validation accuracy (0.9328) but the highest test accuracy (0.9371). It can complement Random Forest.

XGBClassifier: High cross-validation accuracy (0.9439) and comparable test accuracy (0.9348). It adds diversity as another boosting-based algorithm.

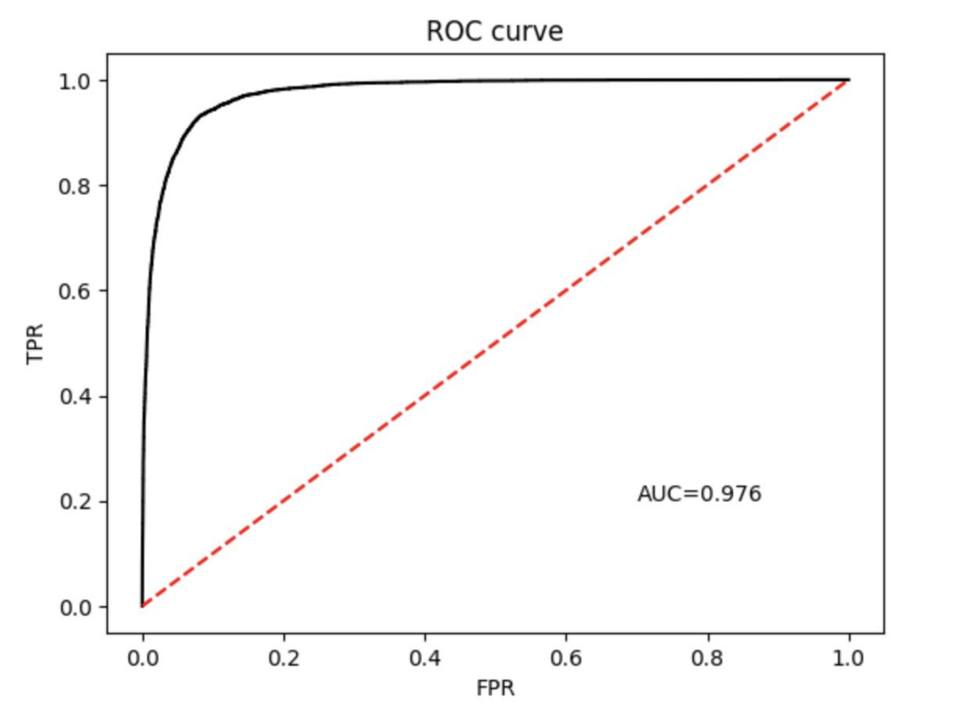


Model	Cross-Validation Accuracy	Test Accuracy	Parameters
RandomForestClassifier	0.9518	0.9356	{'bootstrap': True, 'criterion': 'gini', 'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
XGBClassifier	0.9439	0.9348	{'colsample_bytree': 0.8, 'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 100, 'objective': 'binary:logistic', 'subsample': 1.0}
DecisionTreeClassifier	0.9361	0.9241	{'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5}
CatBoostClassifier	0.9328	0.9371	{'bagging_temperature': 0.5, 'depth': 6, 'iterations': 100, 'l2_leaf_reg': 3, 'learning_rate': 0.05, 'max_bin': 255}
LogisticRegression	0.9320	0.9211	{'C': 0.1, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}
KNeighborsClassifier	0.9113	0.8508	{'algorithm': 'auto', 'n_neighbors': 5, 'weights': 'uniform'}

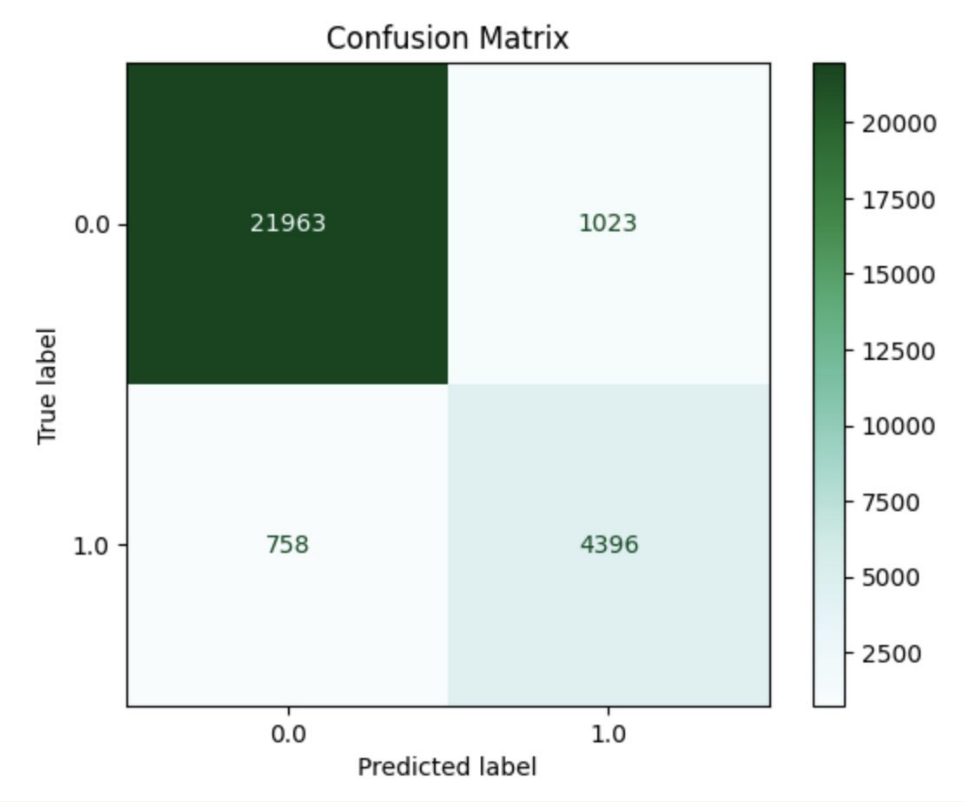
The final model is VotingClassifier ensemble of RandomForestClassifier, XGBClassifier, CatBoostClassifier. The model achieved **0.944** on training set and **0.937** on the test set.



AUC-ROC analysis revealed a score of **0.976**, indicating good separation between classes.



Confusion matrix showed that the model performed well on the majority class but struggled with minority class predictions due to imbalanced data.



Suggestions for Improvement

1. Optimizing the feature engineering process by selecting the most relevant features could lead to better predictive performance.
2. Handle false positive and false negative results.
3. Use feature important analysis for models like XGBoost and CatBoost to discard irrelevant or redundant features.
4. Use additional evaluation techniques to provide insights how to improve model performance.