THE HONG KONG POLYTECHNIC UNIVERSITY

COMP5121 Data Mining and Data Warehousing
Report on Data Mining Practice: Knowledge Discovery from Data

***Abstract*—This study investigates the influence of diverse personal circumstances on depressive mood and clinical depression among a cohort of over 27,000 students. Employing data mining techniques via Python, we analyzed the dataset to identify associations between various factors—including age, academic performance, and perceived academic pressure, etc.—and the prevalence of depression. Preprocessing methodologies were applied to ensure data integrity, followed by the computation of support, lift, and other related metrics to quantify the strength of these relationships. Furthermore, outlier detection was conducted to characterize unique respondent profiles and their potential impact on the overall findings. The resultant conclusions offer empirical insights into the complex interplay of student demographics and psychological well-being, providing a valuable resource for the development of targeted interventions and preventative strategies for student depression.**

***Index Terms*—Academic performance, Data mining, Depression, Outlier detection, Psychological well-being**

I. INTRODUCTION

THE mental health of students has emerged as a critical public health concern, with depressive disorders exhibiting a particularly alarming prevalence. In contemporary academic environments, the confluence of heightened academic pressures, social complexities, and evolving personal circumstances has contributed to a notable escalation in student depression. Recent studies underscore the severity of this issue, with data indicating that a substantial proportion of college students experience significant psychological distress. For example, reports indicate that over three-quarters of college students have reported experiencing moderate to serious psychological distress. Furthermore, approximately 28% of college students have been diagnosed with depression. These statistics highlight the urgent need for comprehensive research aimed at identifying the multifaceted factors that contribute to this pervasive issue.

The impact of student depression extends beyond individual well-being, affecting academic performance, social integration, and long-term life outcomes. The intricate interplay between variables such as age, academic performance, and perceived academic pressure warrants thorough investigation to elucidate the underlying mechanisms that precipitate depressive symptoms. This study aims to address this critical knowledge gap by analyzing a large-scale dataset encompassing over 27,000 respondents. By employing data mining techniques, we seek to quantify the associations between these key variables and the likelihood of depression, thereby providing a robust empirical foundation for understanding the contextual factors that influence student mental health.

Moreover, the incorporation of outlier detection methodologies will enable us to identify and analyze unique respondent profiles, offering nuanced insights that transcend typical population trends. This approach will facilitate a deeper understanding of the diverse experiences within the student population, thereby enhancing the precision of our findings. Ultimately, this research endeavors to provide empirical evidence that can inform the development of targeted interventions and preventative strategies, contributing to the establishment of more effective student mental health support systems. The findings of this study, it is hoped, will provide valuable information for educators, mental health professionals, and policy makers.

The dataset is extracted from Kaggle[1] as a csv file, which includes 27901 rows and 18 columns. In detail, columns 'id', 'Gender', 'Age', 'City', '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', and 'Depression' are included, and each row indicates the information of a single respondent.

II. PREPROCESSING

Prior to conducting substantive analyses, a rigorous preprocessing stage was implemented to ensure data integrity and validity. Initial examination revealed the absence of null values and duplicate records within the dataset. However, anomalies were identified in the 'City' and 'Financial Stress' variables. Specifically, the 'City' column exhibited erroneous entries, including non-categorical values such as ‘M.Com’ and ‘3.0’, which deviate from the expected city names. Similarly, the 'Financial Stress' column contained instances of ‘?’, indicating missing or uninterpretable data. To mitigate the potential bias introduced by these inconsistencies, records containing these errors were systematically removed from the dataset. This cleaning procedure resulted in a refined dataset

1. Adil Shamim, ed., “Student Depression Dataset,” Kaggle, March 13, 2025, https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset.

comprising 27,872 valid observations, which subsequently formed the basis for all subsequent analyses. Note that column "Depression" owns a binary indicator 0/1 (or Yes/No) that denotes whether a student is experiencing depression.

III. Data Visualizations

To facilitate a comprehensive understanding of the dataset's distributional properties, a series of univariate visualizations were generated. Specifically, categorical variables were represented using bar plots, with each category plotted along the x-axis and the corresponding frequency of observations along the y-axis. This approach enabled the clear depiction of the distribution of respondents across various demographic and contextual factors. For instance, Figure 1 illustrates the gender distribution of the respondent cohort, presenting the absolute frequency of male and female participants. Similarly, Figure 2 visualizes the geographical distribution of respondents, showcasing the frequency of observations across different cities. These visualizations provided a foundational understanding of the dataset's composition, informing subsequent analytical procedures. More plots are shown below.

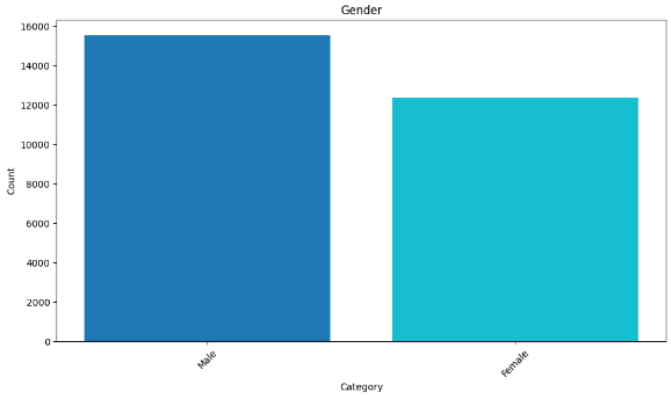

Fig. 1

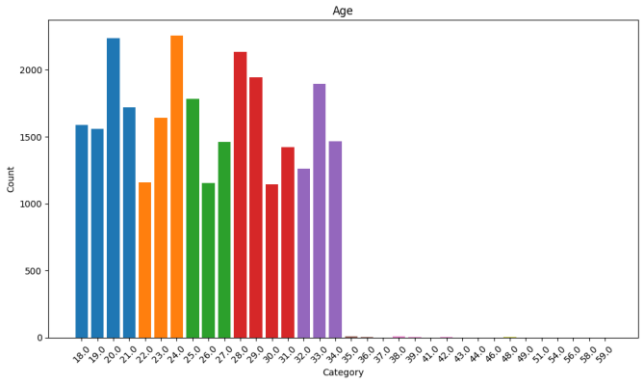

Fig. 2

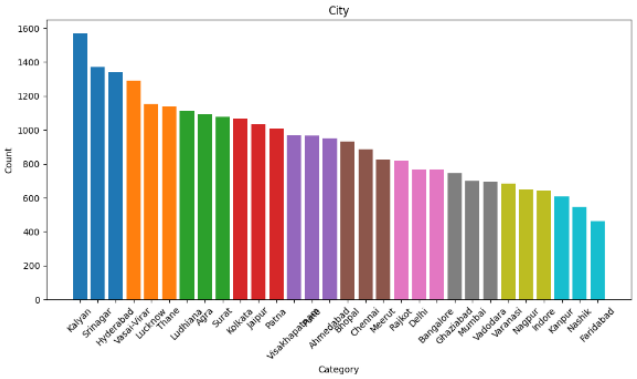

Fig. 3

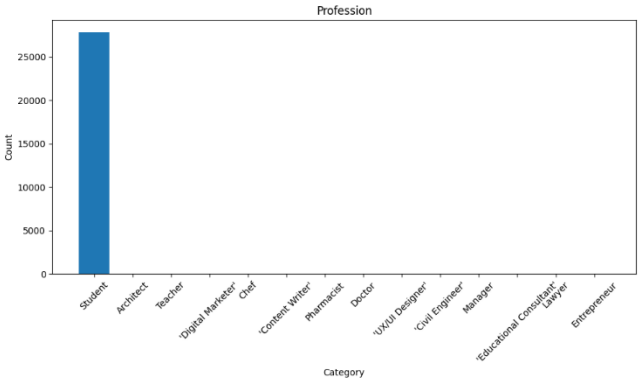

Fig. 4

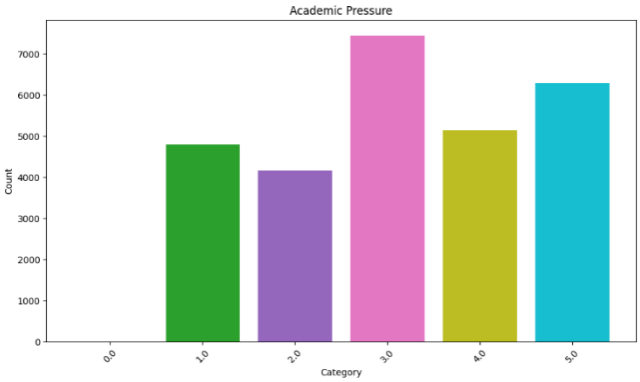

Fig. 5

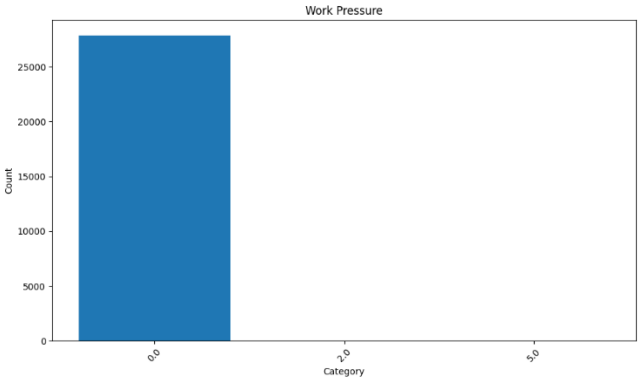

Fig. 6

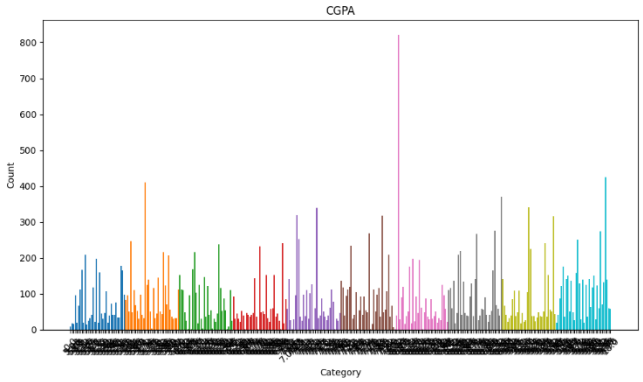

Fig. 7

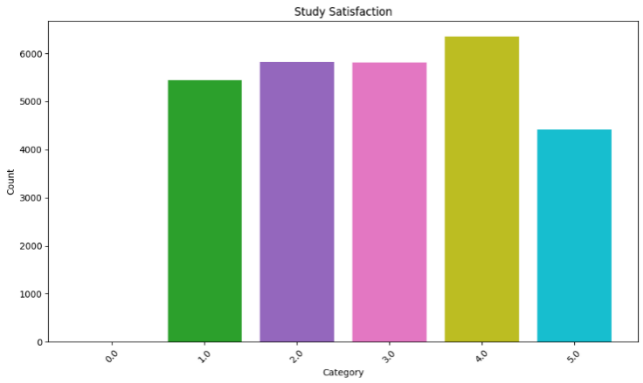

Fig. 8

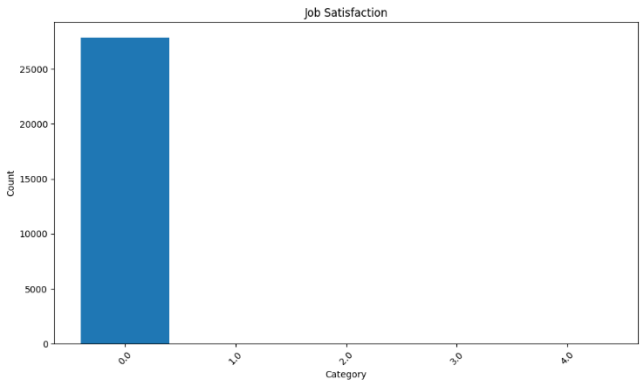

Fig. 9

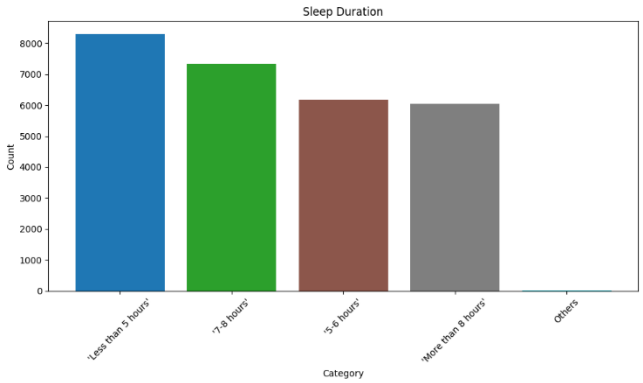

Fig. 10

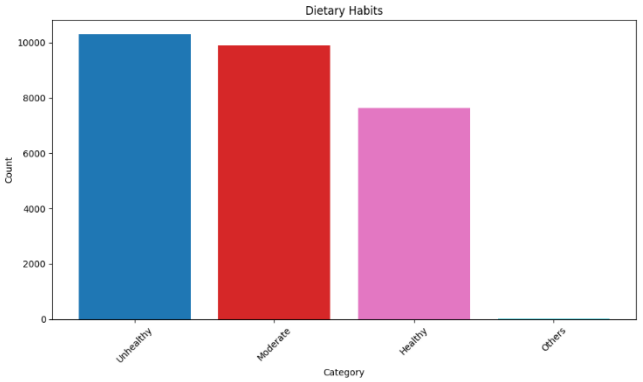

Fig. 11

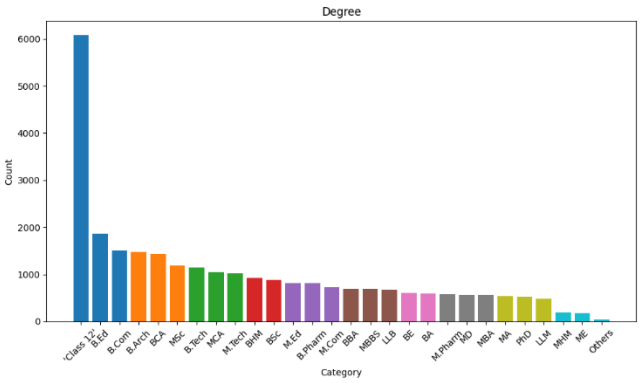

Fig. 12

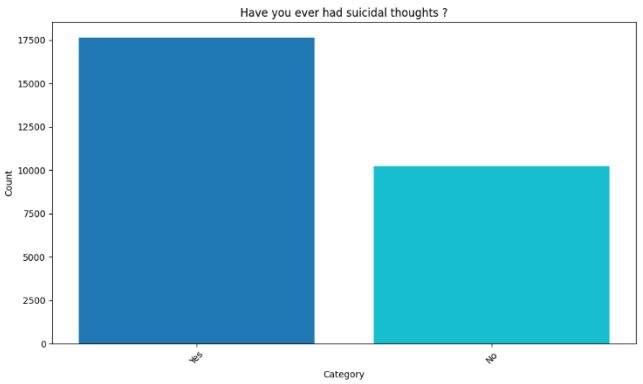

Fig. 13

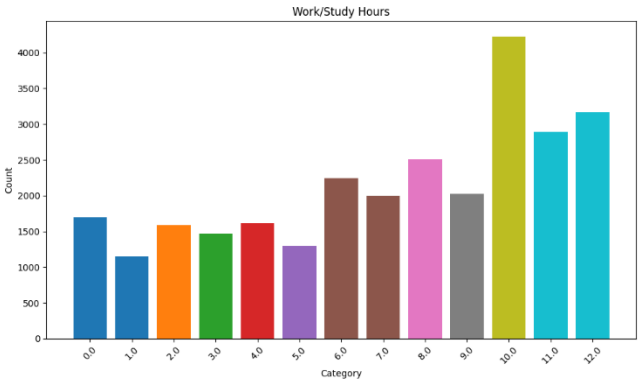

Fig. 14

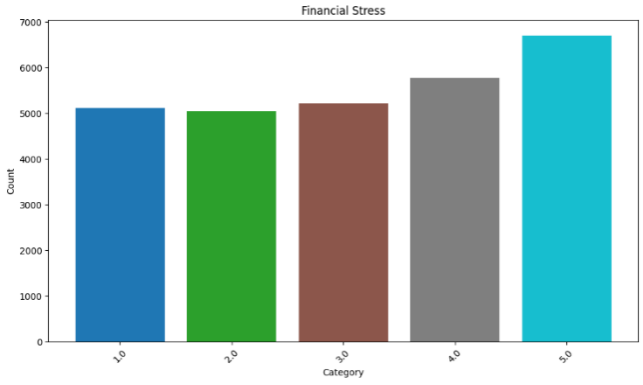

Fig. 15

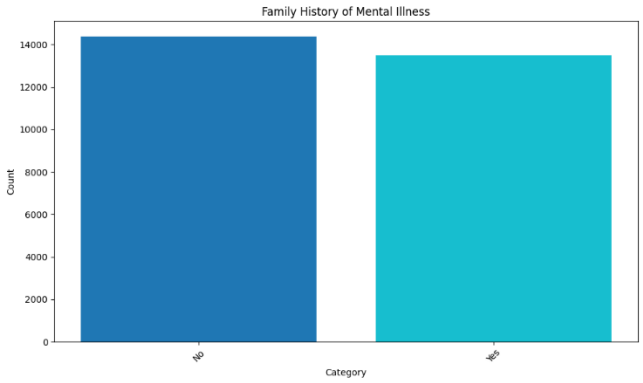

Fig. 16

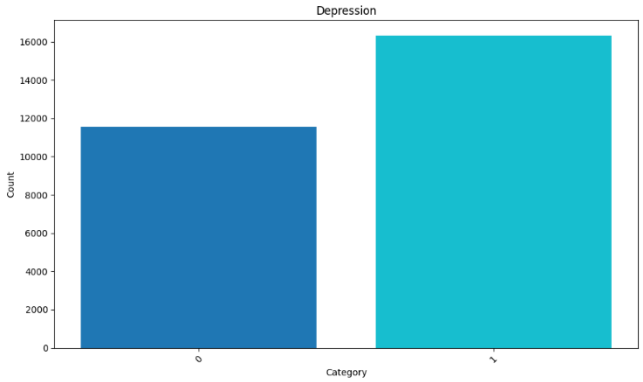

Fig. 17

IV. Outlier Detection

To identify and analyze unique respondent profiles, we performed outlier detection. For example, in 'Academic Pressure' category, only 9 people own 0 pressure. In 'Study Satisfaction' category, only 10 people own 0 satisfaction.

Fig. 18 shows the distribution of academic stress among individuals of different ages and their relationship with depressive condition. We find elderly respondents with high academic pressure tend to have depression compared to youngers with high academic pressure.

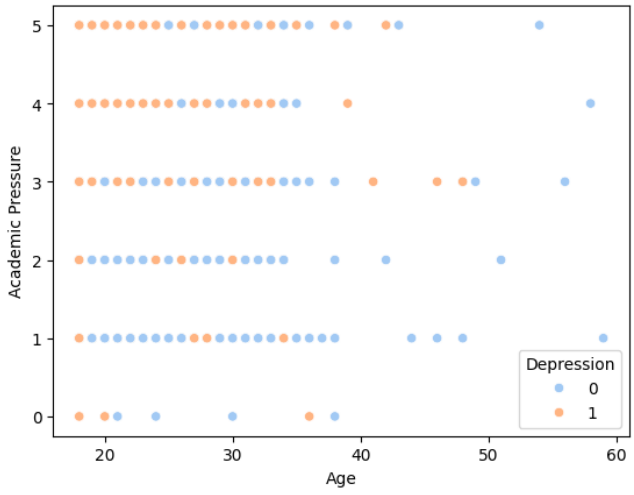

Fig. 18

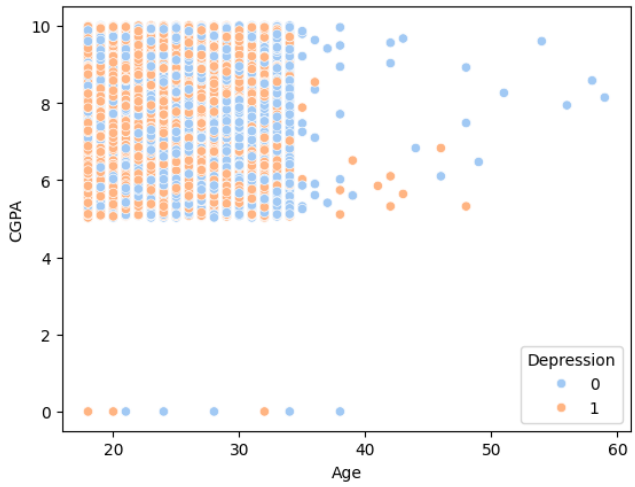

Fig. 19

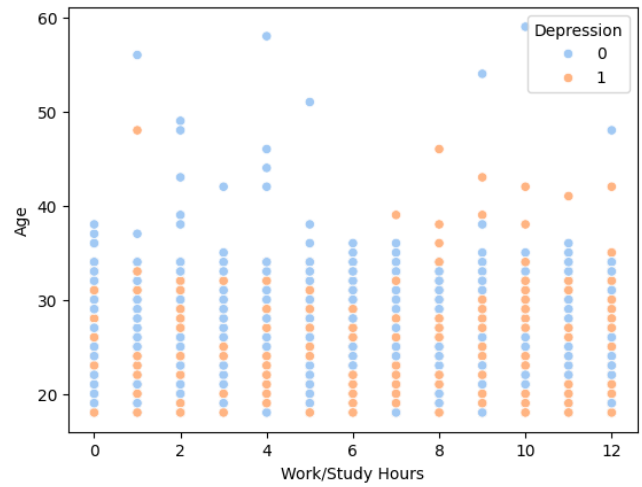

Fig. 20

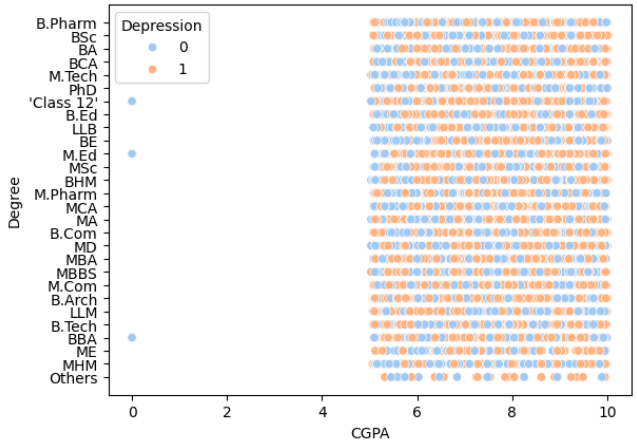

Fig. 21

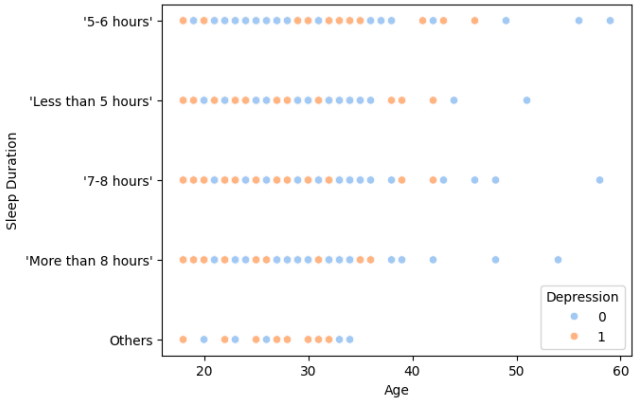

Fig. 22

With the above plots, we find that elder respondents tend to have depression, even though their proportion of all respondents is very small. From outlier detection, we could assume that elderly individuals are more likely to have depression, no matter how their study or life conditions are. However, as the number of elderly people is particularly small, we can easily be misled by chance. In fact, it is not normal to have elderly respondents in a depression survey for students, so the elderly who participated in the survey are likely to suffer from depression, otherwise they might not have participated in the survey at all.

V. Association Rule Mining

To uncover latent relationships between variables and depression, association rule mining was employed using Apriori and FP - Growth algorithms, combined with parameter tuning experiments, to identify key predictive patterns.

A. Data Preparation

Discretization process: Age and Cumulative Grade Point Average (CGPA) were transformed into ordinal categories through standardized binning procedures. Age intervals were operationalized as Young (0-20 years), Adult (21-30), Middle-aged (31-40), and Senior (>40). Academic performance (CGPA) was stratified into Low (0-6.0), Moderate (6.1-8.0), and High (8.1-10.0) tiers.

Label encoding: The dichotomous target variable Depression was recoded into clinically meaningful categories (No Depression vs. Clinical Depression) using standardized nomenclature.

Irrelevant Column Removal: Columns like “id” and “City” unrelated to research objectives were deleted.

B. Association Rule Mining

Algorithm benchmarking: A comparative analysis was conducted between Apriori and FP-Growth algorithms across heterogeneous parameter configurations to evaluate computational efficiency and pattern discovery capability.

Parameter Tuning: Test combinations of minimum support (0.05, 0.1, 0.15) and minimum confidence (0.5, 0.7, 0.9) to evaluate rule quantity and quality.

Rule selection criteria: Statistically significant associations (Lift >1.5) were prioritized, with particular emphasis on depression-correlated patterns demonstrating clinical relevance.

C. Results and Visualization

Both algorithms identify 1331 frequent itemsets, demonstrating comparable performance in itemset frequency detection.

	algorithm	min_support	min_confidence	num_rules
0	apriori	0.05	0.5	78729
1	fpgrowth	0.05	0.5	78729
2	apriori	0.05	0.7	33443
3	fpgrowth	0.05	0.7	33443
4	apriori	0.05	0.9	19359
5	fpgrowth	0.05	0.9	19359
6	apriori	0.10	0.5	18514
7	fpgrowth	0.10	0.5	18514
8	apriori	0.10	0.7	7968
9	fpgrowth	0.10	0.7	7968
10	apriori	0.10	0.9	4855
11	fpgrowth	0.10	0.9	4855
12	apriori	0.15	0.5	7008
13	fpgrowth	0.15	0.5	7008
14	apriori	0.15	0.7	3257
15	fpgrowth	0.15	0.7	3257
16	apriori	0.15	0.9	1923
17	fpgrowth	0.15	0.9	1923

Fig. 23

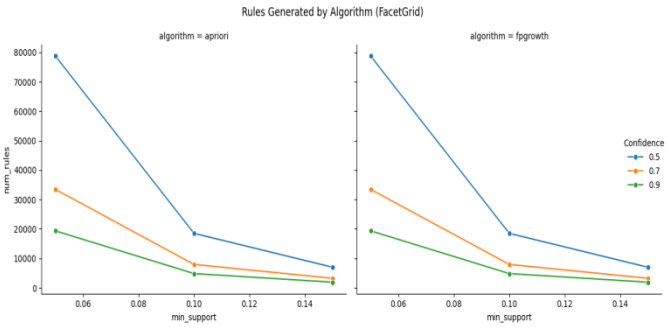

Fig. 24

As shown in the two figures above, under identical min_support and min_confidence conditions, Apriori and FP-Growth generate the same number of rules.For example, with min_support=0.05 and min_confidence=0.5, both algorithms generate 78729 rules.Higher min_support and min_confidence values result in fewer rules.Under stringent thresholds (support=0.15, confidence=0.9), the unified output of 1923 rules achieves optimal balance between rule abundance and clinical interpretability.

Rules were filtered using lift (>1.5) to identify non-random associations.The significant patterns associated with depression are shown in Figure 25.

Top Association Rules with Strong Lift (>1.5):

	antecedents	consequents
465	(Depression, 'Class 12')	(Yes, Young, Student, 0.0)
466	(Yes, Young)	(Student, Depression, 0.0, 'Class 12')
457	(Depression, 0.0, 'Class 12')	(Yes, Young, Student)
458	(Yes, Young, Student)	(Depression, 0.0, 'Class 12')
459	(Yes, Young, 0.0)	(Student, Depression, 'Class 12')
287	(Depression, 0.0, 'Class 12')	(Yes, Young)
292	(Yes, Young)	(Depression, 0.0, 'Class 12')
291	(Depression, 'Class 12')	(Yes, Young, 0.0)
288	(Yes, Young, 0.0)	(Depression, 'Class 12')
355	(Depression, 'Class 12')	(Yes, Young, Student)

	antecedent support	consequent support	support	confidence	lift
465	0.154241	0.162887	0.128229	0.831356	5.103867
466	0.162887	0.154241	0.128229	0.787225	5.103867
457	0.154241	0.162887	0.128229	0.831356	5.103867
458	0.162887	0.154241	0.128229	0.787225	5.103867
459	0.162887	0.154241	0.128229	0.787225	5.103867
287	0.154241	0.162887	0.128229	0.831356	5.103867
292	0.162887	0.154241	0.128229	0.787225	5.103867
291	0.154241	0.162887	0.128229	0.831356	5.103867
288	0.162887	0.154241	0.128229	0.787225	5.103867
355	0.154241	0.162887	0.128229	0.831356	5.103867

Fig. 25

The scatter plot visualizes rules by support, confidence, and lift (size). Rules with higher lift (e.g., 4.8) cluster in areas of moderate support (0.10–0.16) and confidence (0.5–1.0), underscoring their significance.

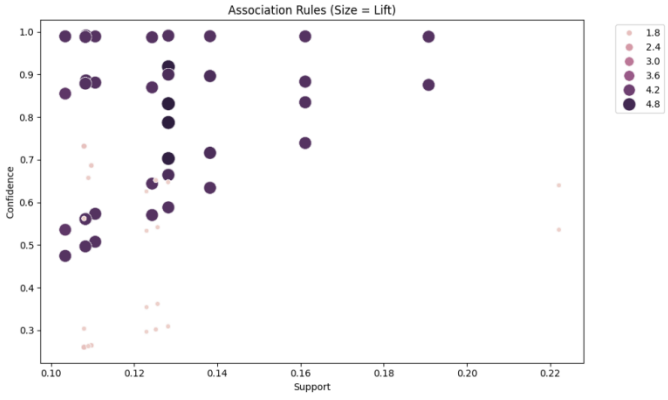

Fig. 26

Filter the top 5 association rules containing 'Depression' from the highest-lift rules,including:

If Young, Yes, Student, then 'Class 12', Depression (confidence: 0.79, lift: 5.10)

If Young, Yes, then 0.0, 'Class 12', Depression, Student (confidence: 0.79, lift: 5.10)

If Young, Yes, then 'Class 12', Depression (confidence: 0.79, lift: 5.10)

If Young, Yes, then 0.0, 'Class 12', Depression (confidence: 0.79, lift: 5.10)

If 0.0, Yes, Young, then 'Class 12', Depression (confidence: 0.79, lift: 5.10)

These rules highlight critical combinations of demographic, academic, and lifestyle factors that strongly predict depression.

For example,rules consistently link young students (aged 0-20) with academic factors (e.g., "Class 12") to depression.

Moreover,all top rules have a lift score of 5.10, indicating these patterns are 5.1 times more likely to occur than random chance.The rules suggest interventions targeting younger students in high-pressure academic environments could reduce depression risk.

VI. Classification

In this section, we used three classification models to predict student depression: Logistic Regression, Random Forest Classifier, and Decision Tree Classifier. Our choice was based on the characteristics of the data, and we evaluated the models using various performance metrics.

A. Data Preparation

Feature Selection: We dropped the id, City, and Depression columns, using Depression as the target variable.

Encoding Categorical Variables: We used LabelEncoder to encode all categorical variables.

Data Standardization: We used StandardScaler to standardize the features.

Data Balancing: Due to the imbalance in the target variable, we used the SMOTE technique to balance the dataset.

Data Splitting: The data was split into 80% training set and 20% test set.

B. Model Building and Evaluation

Logistic Regression

The Logistic Regression model shows a good balance in predicting student depression, with precision, recall, and F1 scores all at 0.84. This indicates that the model performs similarly in identifying both depressed and non-depressed students.

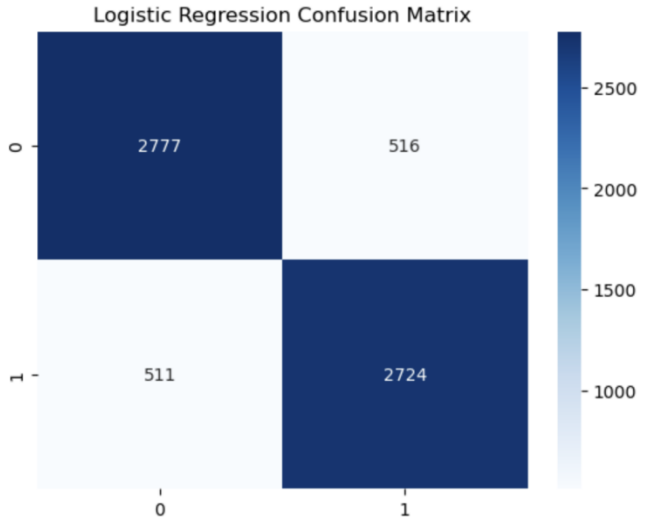

Fig. 27

	precision	recall	f1-score	support
0	0.84	0.84	0.84	3293
1	0.84	0.84	0.84	3235
accuracy			0.84	6528
macro avg	0.84	0.84	0.84	6528
weighted avg	0.84	0.84	0.84	6528

Tab. 1

Random Forest Classifier

The Random Forest Classifier performs better across all metrics, with an accuracy of 0.8632 and precision, recall, and F1 scores all at 0.86. This indicates that the Random Forest is

more consistent and accurate in identifying both depressed and non-depressed students.

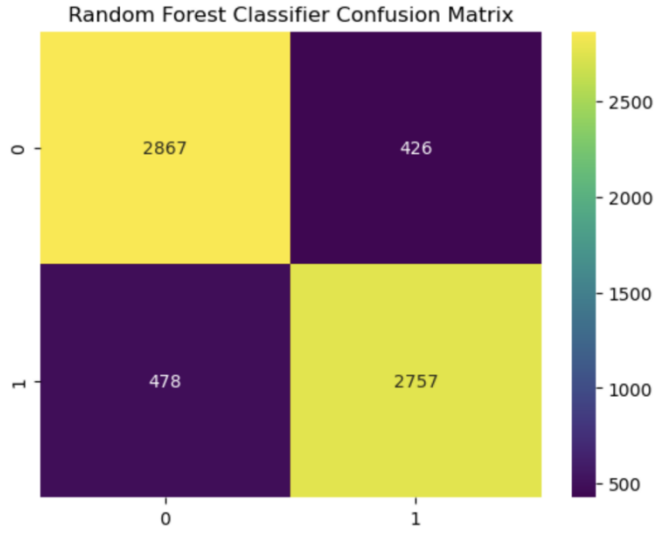

Fig. 28

	precision	recall	f1-score	support
0	0.86	0.87	0.86	3293
1	0.86	0.86	0.86	3235
accuracy			0.86	6528
macro avg	0.86	0.86	0.86	6528
weighted avg	0.86	0.86	0.86	6528

Tab. 2

Decision Tree Classifier

The Decision Tree Classifier has an accuracy of 0.8208, with precision and recall of 0.85 and 0.78 for class 0, and 0.79 and 0.86 for class 1. This indicates that the Decision Tree performs better at identifying non-depressed students but slightly worse at identifying depressed students.

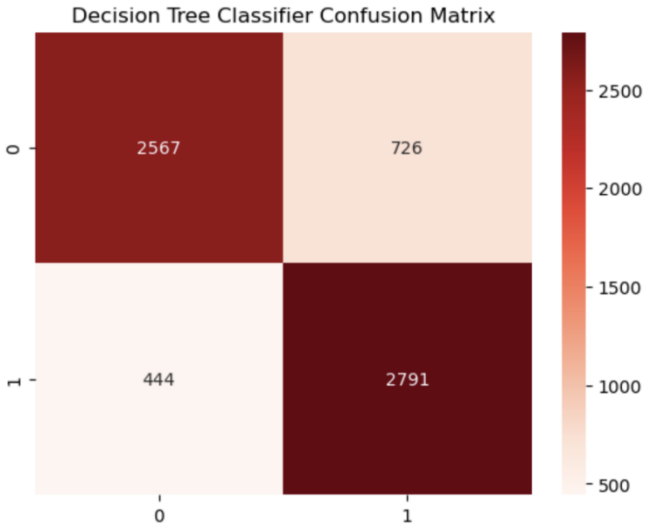

Fig. 29

	precision	recall	f1-score	support
0	0.85	0.78	0.81	3293
1	0.79	0.86	0.83	3235
accuracy			0.82	6528
macro avg	0.82	0.82	0.82	6528
weighted avg	0.82	0.82	0.82	6528

Tab. 3

We employed 5-fold cross-validation to assess and compare the performance of our models, which included Logistic Regression, Random Forest Classifier, and Decision Tree Classifier. The results indicated that Logistic Regression achieved an average accuracy of 0.8421, the Random Forest Classifier attained an average accuracy of 0.8627, and the Decision Tree Classifier recorded an average accuracy of 0.8205. These findings from the cross-validation process further substantiate that the Random Forest Classifier outperforms the other models, with Logistic Regression coming in second, while the Decision Tree Classifier shows comparatively lower performance.

The superior performance of the Random Forest Classifier can be attributed to its ability to handle complex, non-linear relationships within the data through the construction of multiple decision trees and the aggregation of their outputs. This ensemble approach not only reduces variance but also enhances the model's predictive accuracy. On the other hand, Logistic Regression, while simpler and more interpretable, still provides robust results, making it a valuable tool for understanding the impact of various predictors on the likelihood of depression among students. The Decision Tree Classifier, despite its intuitive decision-making process, tends to overfit the training data, which may explain its lower accuracy in our cross-validation results. By carefully analyzing these outcomes, we can better tailor our approach to predicting and addressing student mental health issues.

We chose Logistic Regression, Random Forest Classifier, and Decision Tree Classifier for this classification task based on their unique strengths. Logistic Regression is a simple and interpretable model that can handle both categorical and numerical data. The coefficients of Logistic Regression can help us understand the impact of each feature on the prediction. For example, features with larger coefficients have a greater impact on the prediction. Random Forest Classifier, as an ensemble method, can capture complex relationships in the data and is less prone to overfitting. It makes predictions by voting among multiple decision trees, which makes it excellent for handling high-dimensional data and non-linear relationships. Decision Tree Classifier provides a clear visualization of the decision-making process and can be easily interpreted. The rules of the Decision Tree can be directly extracted from its structure, helping us understand how the model makes predictions.

To ensure the robustness of our analysis, we chose to split the data into 80% training and 20% testing sets, and used SMOTE to balance the dataset due to the imbalance in the target variable. Model performance was evaluated using accuracy, precision, recall, and F1 score. The confusion matrices provide a visual representation of each model's performance. Cross-validation was used to get a more robust estimate of model performance, which confirmed the Random

Forest Classifier as the top performer, followed by Logistic Regression, and then the Decision Tree Classifier. This comprehensive approach allows us to make informed decisions about the best model to use for predicting and addressing student mental health issues.

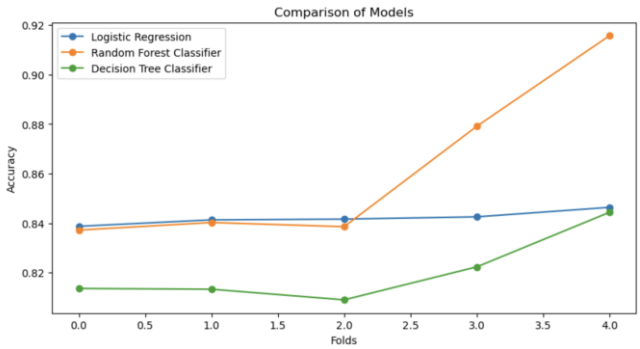

Fig. 30

C. Decision Tree Rules Analysis

From the Decision Tree rules, we can see that the model primarily relies on the following three features for prediction:

1. Suicidal Thoughts (Have you ever had suicidal thoughts?): This is the most important feature, which divides the data into two main categories based on its value.

2. Academic Pressure: In cases where suicidal thoughts are low, academic pressure further segments the data.

3. Financial Stress: In cases where academic pressure is high or suicidal thoughts are high, financial stress further influences the prediction.

For example, if a student has no suicidal thoughts (suicidal thoughts <= 0.53) and low academic pressure (academic pressure <= 0.62), the model predicts that the student is not depressed (class 0), regardless of financial stress. Conversely, if a student has suicidal thoughts (suicidal thoughts > 0.53) and high academic pressure (academic pressure > -0.11), the model predicts that the student is depressed (class 1), regardless of financial stress.

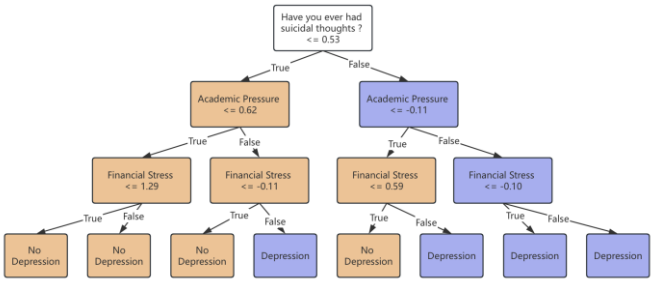

Fig. 31

D. Feature Importance Analysis

From the feature importance analysis, it is evident that suicidal thoughts, academic pressure, and financial stress are the three most important factors in predicting student depression. This aligns with the observations from the Decision Tree rules, further validating the significance of these features in the models.

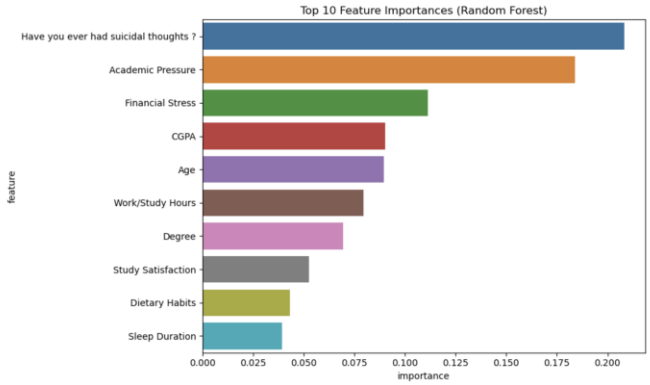

Fig. 32

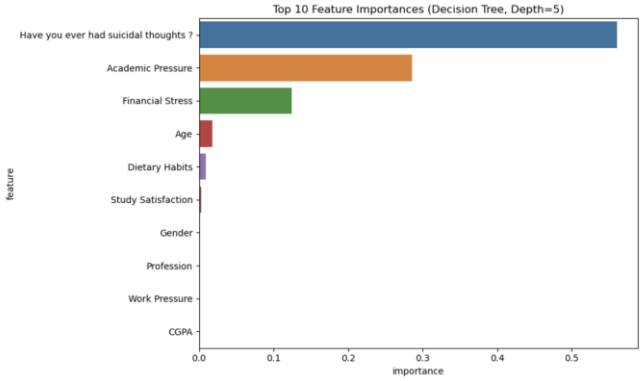

Fig. 33

E. Model Output Interpretation

Logistic Regression: By examining the coefficients of the Logistic Regression model, we can understand the impact of each feature on the prediction. For example, if the coefficient for suicidal thoughts is positive and large, it indicates that suicidal thoughts have a significant positive impact on predicting depression.

Random Forest: The feature importance of the Random Forest can help us identify which features contribute the most to the prediction. For instance, suicidal thoughts and academic pressure are the two most important features, indicating that these factors play a crucial role in predicting student depression.

Decision Tree: The visualization and rules of the Decision Tree can help us understand how the model uses these features to make predictions. For example, if a student has suicidal thoughts and high academic pressure, the model will predict that the student is depressed.

F. Model Application and Extension

Early Intervention: By identifying key factors such as suicidal thoughts, academic pressure, and financial stress, schools and mental health professionals can implement early interventions to help students alleviate these stresses and prevent the onset of depression.

Personalized Interventions: Based on each student's specific situation, personalized intervention plans can be developed. For example, for students with high academic pressure, academic tutoring and stress management courses can be provided.

Model Improvement: We can consider using more features or more complex models (such as neural networks) to further improve prediction accuracy. Additionally, collecting more data to train the models can enhance their generalization capabilities.

Real-time Monitoring: Developing a real-time monitoring system to periodically assess students' depression risk and issue timely alerts when the risk increases.

Through this approach, we can not only predict the risk of student depression but also provide targeted interventions to help students better manage their mental health.

REFERENCES

[1] Ibrahim, A. K., Kelly, S. J., Adams, C. E., & Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. *Journal of psychiatric research*, *47*(3), 391-400.

[2] Hysenbegasi, A., Hass, S. L., & Rowland, C. R. (2005). The impact of depression on the academic productivity of university students. *Journal of mental health policy and economics*, *8*(3), 145.

[3] Sarokhani, D., Delpisheh, A., Veisani, Y., Sarokhani, M. T., Esmaeli Manesh, R., & Sayehmiri, K. (2013). Prevalence of depression among university students: A systematic review and meta-analysis study. *Depression research and treatment*, *2013*(1), 373857.

[4] Romaniuk, M., & Khawaja, N. G. (2013). University student depression inventory (USDI): Confirmatory factor analysis and review of psychometric properties. *Journal of affective disorders*, *150*(3), 766-775.

[5] Silva, V., Costa, P., Pereira, I., Faria, R., Salgueira, A. P., Costa, M. J., ... & Morgado, P. (2017). Depression in medical students: insights from a longitudinal study. *BMC medical education*, *17*, 1-9.

[6] Sharif, A. R., Ghazi-Tabatabaei, M., Hejazi, E., Askarabad, M. H., & Dehshiri, G. R. (2011). Confirmatory factor analysis of the university student depression inventory (USDI). *Procedia-Social and Behavioral Sciences*, *30*, 4-9.

[7] Knudson-Martin, C., & Silverstein, R. (2009). Suffering in Silence: A qualitative meta-data-analysis of postpartum depression. *Journal of marital and family therapy*, *35*(2), 145-158.

[8] Rapaport, M. H., Judd, L. L., Schettler, P. J., Yonkers, K. A., Thase, M. E., Kupfer, D. J., ... & Rush, A. J. (2002). A descriptive analysis of minor depression. *American Journal of Psychiatry*, *159*(4), 637-643.

[9] Jadoon, N. A., Yaqoob, R., Raza, A., Shehzad, M. A., & Zeshan, S. C. (2010). Anxiety and depression among medical students: a cross-sectional study. *JPMA. The Journal of the Pakistan Medical Association*, *60*(8), 699-702.

[10] Kleine-Budde, K., Müller, R., Kawohl, W., Bramesfeld, A., Moock, J., & Rössler, W. (2013). The cost of depression–a cost analysis from a large database. *Journal of affective disorders*, *147*(1-3), 137-143.