Counterfactual Analysis for Student Depression Prediction

Course No.: DATA230

Course Name: Data Visualization

M.S Data Analytics

SAN JOSE STATE UNIVERSITY

SAN JOSE, CA

PROJECT REPORT

GROUP-3:

Aakash Vardhan Madabhushi (018291663)

Aishwarya Iyer (018277948)

Keith Rajesh Gonsalves (018326945)

Om Dankhara (018221463)

May 2025

Acknowledgment

We would like to express our deepest gratitude to all those who supported and guided us throughout the course of this project.

Our sincere thanks go to Prof. Guannan Liu for her exceptional mentorship, insightful feedback, and constant encouragement. Her expertise and support were instrumental in shaping the direction and success of this work.

We are also thankful to the Department of Applied Data Science for providing us with the academic resources, infrastructure, and collaborative environment that made this research possible.

We are especially grateful to our families and friends for their unwavering support, patience, and motivation during this journey. Their encouragement helped us stay focused and committed through every phase of the project.

This project has been a significant learning experience, offering us the opportunity to apply theoretical knowledge to a real-world problem while growing both individually and as a team. We are proud of what we have accomplished together and thankful for the opportunity to collaborate on work that aims to make a meaningful impact.

Table of Contents

Figure Table

Abstract

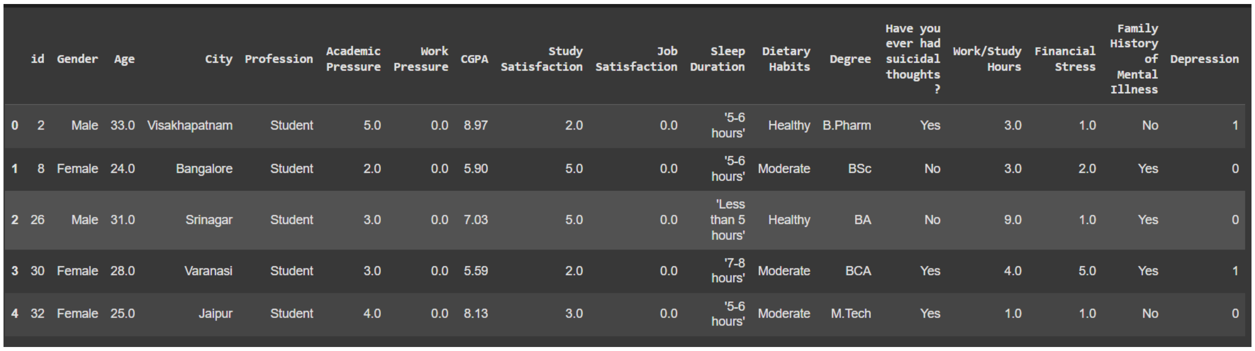
This project focuses on predicting student depression using interpretable machine learning techniques and generating actionable counterfactuals. The goal is to not only identify students at risk of depression but also offer feasible recommendations for intervention. The dataset used contains demographic, academic, behavioral, and psychological information of students. Multiple models were trained including Random Forest and Logistic Regression, with and without Synthetic Minority Oversampling Technique (SMOTE) for class balancing. Logistic Regression (without SMOTE) was chosen for its balance of interpretability and performance. We used logistic regression coefficients for feature importance and DiCE (Diverse Counterfactual Explanations) to generate realistic interventions for students predicted to be depressed.

1. Introduction

Student depression is becoming a growing concern due to academic pressure, lifestyle imbalance, and socio-environmental causes. While traditional ML models can classify at-risk students, they are less interpretable and cannot provide recommendations to reduce the risk. Recent research emphasizes explainable AI in sensitive domains. Logistic Regression, while basic, offers transparency via coefficients. DiCE counterfactuals are emerging tools that recommend minimal, realistic feature changes to flip an undesired prediction, offering actionable insights. This report explores both prediction and intervention using these approaches.

2. Dataset

The dataset includes over 27,900 student records, featuring demographic (Age, Gender, City, Profession), academic (Academic Pressure, CGPA), behavioral (Dietary Habits, Work/Study Hours), and psychological (Suicidal Thoughts, Financial Stress) attributes. The target variable is binary: 'Depression' (1 for depressed, 0 for not).



	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	Depression
0	2	Male	33.0	Visakhapatnam	Student	5.0	0.0	8.97	2.0	0.0	'5-6 hours'	Healthy	B.Pharm	Yes	3.0	1.0	No	1
1	8	Female	24.0	Bangalore	Student	2.0	0.0	5.90	5.0	0.0	'5-6 hours'	Moderate	BSc	No	3.0	2.0	Yes	0
2	26	Male	31.0	Srinagar	Student	3.0	0.0	7.03	5.0	0.0	'Less than 5 hours'	Healthy	BA	No	9.0	1.0	Yes	0
3	30	Female	28.0	Varanasi	Student	3.0	0.0	5.59	2.0	0.0	'7-8 hours'	Moderate	BCA	Yes	4.0	5.0	Yes	1
4	32	Female	25.0	Jaipur	Student	4.0	0.0	8.13	3.0	0.0	'5-6 hours'	Moderate	M.Tech	Yes	1.0	1.0	No	0

Fig 1: Dataset overview

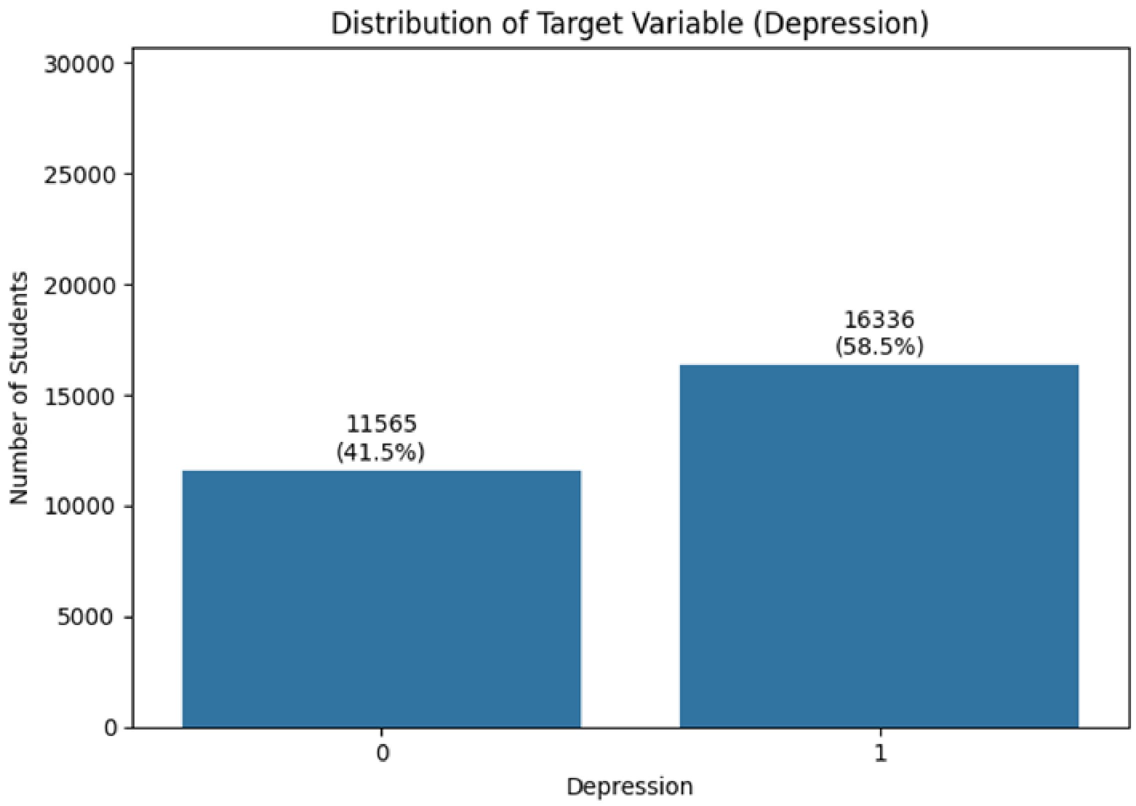


Figure 2: Distribution of Depression; The target variable is imbalanced with 58.5% of students labeled as depressed.

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

3. Data Exploration

Exploratory Data Analysis revealed key patterns:
- Academic Pressure and Suicidal Thoughts are strongly associated with depression
- Financial Stress and Study Satisfaction also play a major role
- Several features are skewed or unevenly distributed

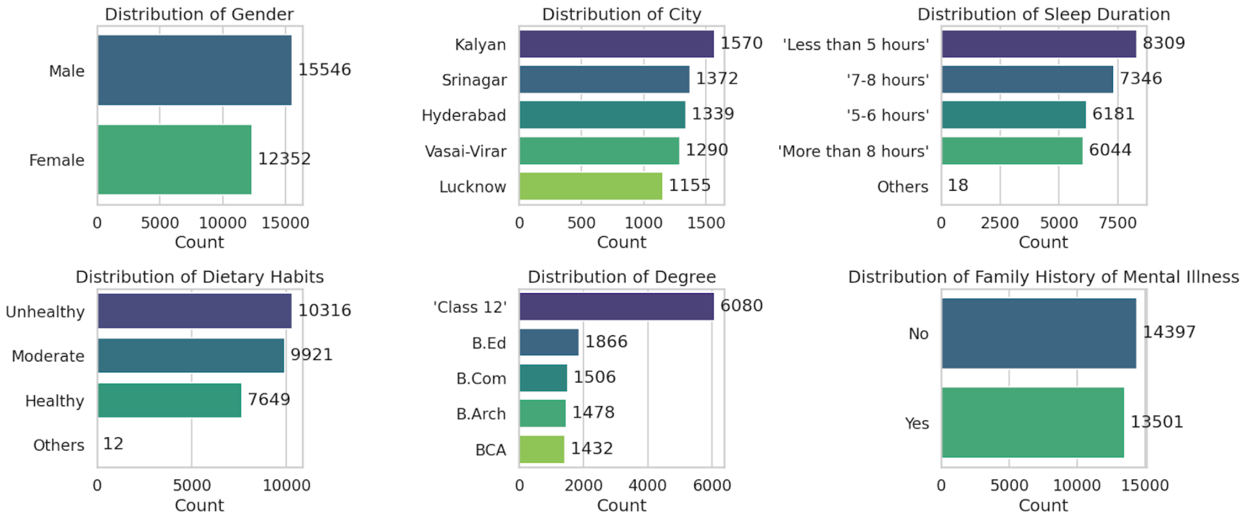


Fig 3: Categorical Univariant Analysis

Figure 3 presents the frequency distribution for each categorical variable in the dataset. The key observations are summarized below:

Gender

- 15546 males (56 %) versus 12352 females (44 %)
- A moderate male skew that should be controlled for in downstream analyses

City

- Top five cities are Kalyan (1570), Srinagar (1372), Hyderabad (1339), Vasai-Virar (1290) and Lucknow (1155)
- No single city dominates, but the sample is geographically clustered around these locations

Sleep Duration

- "Less than 5 hours" is most common (8309), followed by "7–8 hours" (7346) and "5–6 hours" (6181)
- Over 50 % of students obtain fewer than 7 hours per night—an important consideration for mental-health modeling

Dietary Habits

- "Unhealthy" (10316) and "Moderate" (9921) both substantially exceed "Healthy" (7649)
- Indicates a majority with at-risk eating patterns

Degree Level

- Nearly half of the cohort are in “Class 12” (6080)
- Undergraduates form much smaller groups
    - B.Ed: 1866 samples
    - B.Com: 1506 samples
    - B.Arch: 1478 samples
    - BCA: 1432 samples

Family History of Mental Illness

- “No” family history (14397) slightly outweighs “Yes” (13501)
- Roughly balanced split—family history may be a useful predictor but not overwhelmingly dominant

Implication:

The sample is skewed toward male, pre-college students from a handful of cities, with generally poor sleep and dietary habits. Failure to adjust for these imbalances risks overfitting models to this demographic segment.

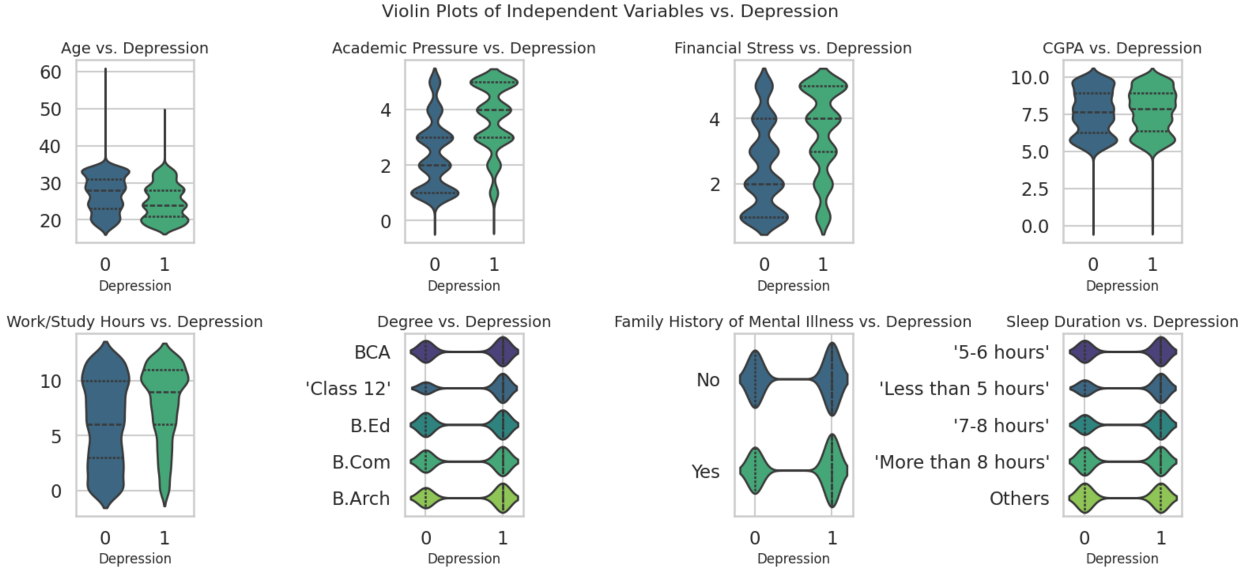


Fig 4: Violin Plots of Independent Variables vs. Depression

Figure 4 compares the distribution of each predictor for non-depressed (0) and depressed (1) respondents. Key takeaways are as follows:

Age

- Both groups cluster in their mid-20s (medians ≈ 27 years).
- Non-depressed tails extend up to ~60 years vs. ~50 years for people with depression risk.

Inference: age shows no meaningful separation.

Academic Pressure

- Non-depressed: median ≈ 2.5 (on 0–5 scale); depressed: median ≈ 4.0.
- Depressed density concentrated at high-pressure scores (3–5).

Inference: strong positive association with depression.

Financial Stress

- Non-depressed median ≈ 2; depressed ≈ 4.
- Depressed respondents cluster at higher stress levels (3–5).

Inference: mirrors academic-pressure effect.

CGPA

- Non-depressed: median ≈ 8.0; depressed: ≈ 7.5.
- Substantial overlap in distributions.

Inference: CGPA alone is a weak discriminator.

Work/Study Hours

- Non-depressed median ≈ 10 h/day; depressed ≈ 11 h/day.
- Depressed group exhibits a heavier upper tail (up to ~13 h).

Inference: slight tendency toward overwork among depressed students.

Degree Level

- Distributions across Class 12, B.Ed, B.Com, B.Arch, BCA overlap heavily for both classes.

Inference: degree program has minimal impact on depression status.

Family History of Mental Illness

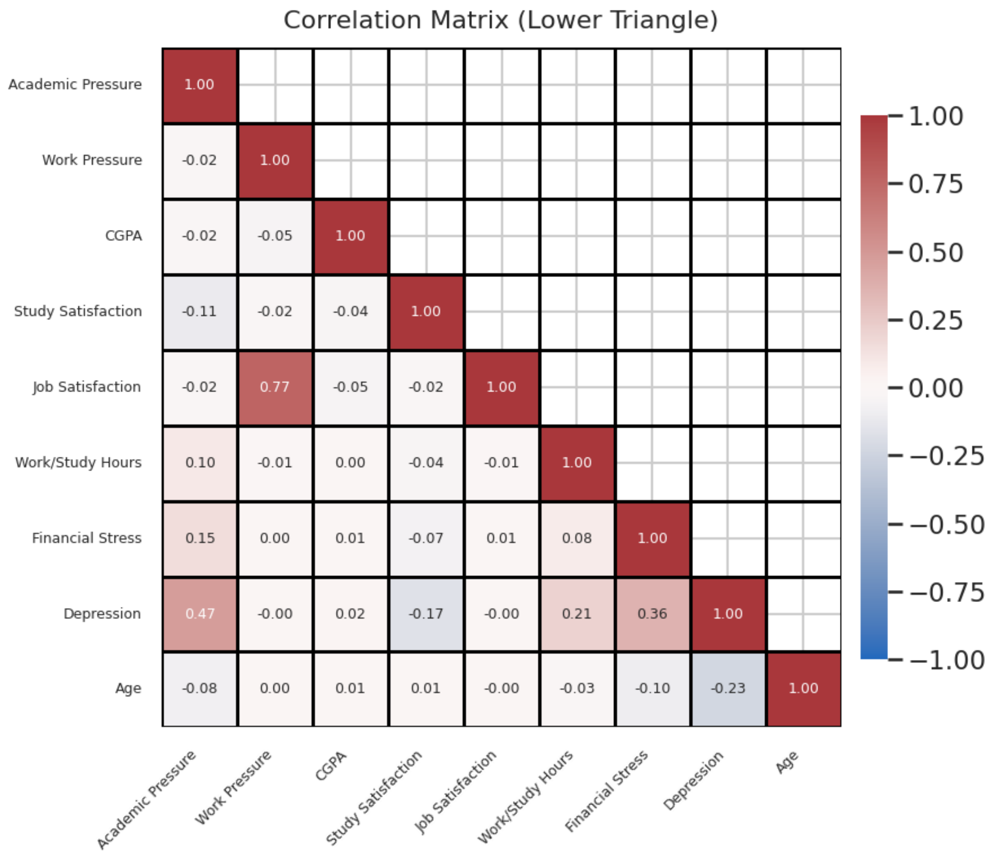
- "Yes" group shows a modestly higher concentration at depression=1 than "No.

Inference: familial predisposition contributes to risk.

Sleep Duration

- Depressed individuals are overrepresented in “< 5 h” and “5–6 h” sleepers; non-depressed dominate “7–8 h” and “> 8 h.”

Inference: short sleep duration is a significant correlate.



	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Work/Study Hours	Financial Stress	Depression	Age
Academic Pressure	1.00								
Work Pressure	-0.02	1.00							
CGPA	-0.02	-0.05	1.00						
Study Satisfaction	-0.11	-0.02	-0.04	1.00					
Job Satisfaction	-0.02	0.77	-0.05	-0.02	1.00				
Work/Study Hours	0.10	-0.01	0.00	-0.04	-0.01	1.00			
Financial Stress	0.15	0.00	0.01	-0.07	0.01	0.08	1.00		
Depression	0.47	-0.00	0.02	-0.17	-0.00	0.21	0.36	1.00	
Age	-0.08	0.00	0.01	0.01	-0.00	-0.03	-0.10	-0.23	1.00

Fig 5: Heatmap shows relationships between key numeric features such as Academic Pressure and Depression.

This heatmap displays pairwise Pearson correlations between all numeric variables in the dataset, including binary-coded variable Depression (1 = Yes, 0 = No). Color value and saturation indicate strength and direction of linear relationship:

Academic Pressure shows the strongest **positive correlation** with Depression (**r = 0.47**). → Higher academic stress is clearly associated with higher likelihood of depression.

Financial Stress has a slight **positive correlation** with Depression (**r = 0.36).**

→ Higher Financial Stress = higher depression risk

Age has a **negative correlation** with Depression (**r = –0.23**). → Younger individuals are more prone to depression symptoms in this dataset.

Study Satisfaction also correlates **negatively** with Depression (**r = –0.17**).
→ Students more satisfied with their studies tend to report lower depression levels.

Work/Study Hours is **positively correlated** with Depression (**r = 0.21**).
→ Longer hours may be linked to burnout or mental fatigue.

CGPA, **Job Satisfaction**, and **Work Pressure** all show **very weak or no correlation** with Depression.
→ These features may not directly influence depressive symptoms or may interact in non-linear ways.

A strong **positive correlation** exists between **Job Satisfaction** and **Work Pressure** (**r = 0.77**).
→ This could indicate multicollinearity or overlapping dimensions in perception of work-related stress.

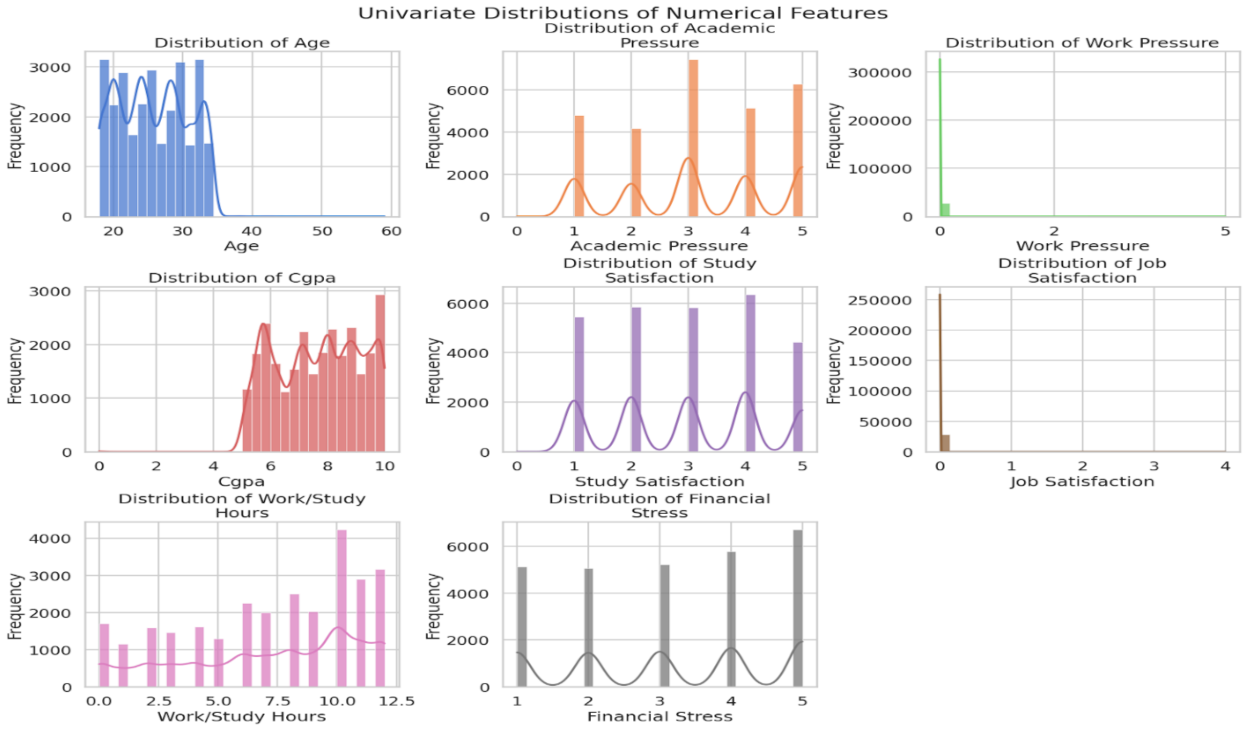


Fig 6: Distribution of Selected Features

4. System Architecture

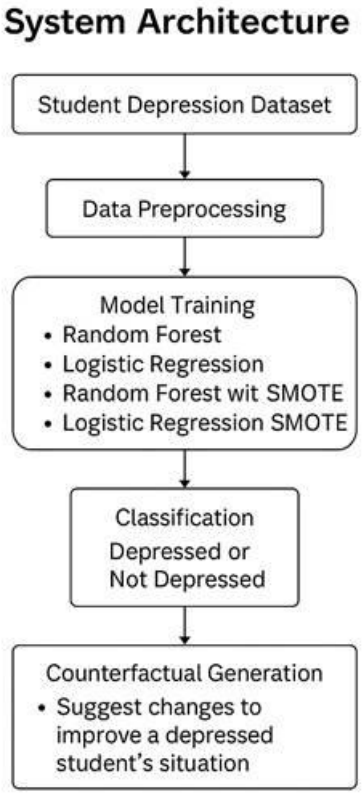


Fig 7: Pipeline from raw data to model training, classification, and counterfactual generation.

The system architecture diagram illustrates the end-to-end process flow of the student depression prediction project. It begins with a massive Student Depression Dataset, which is passed through data preprocessing to clean, encode, and balance features for modeling. Preprocessed data is used to train some models, like Random Forest, Logistic Regression, and their equivalent with SMOTE to handle class imbalance. After training, the models classify, placing each student in either Depressed or Not Depressed. The final step is Counterfactual Generation, where the DiCE library is used to propose small, plausible feature modifications—e.g., reducing academic pressure or improved sleep—that would undo a 'depressed' prediction to 'not depressed'. This architecture presents a structured approach to both identifying at-risk students and providing actionable recommendations for enhancing mental health.

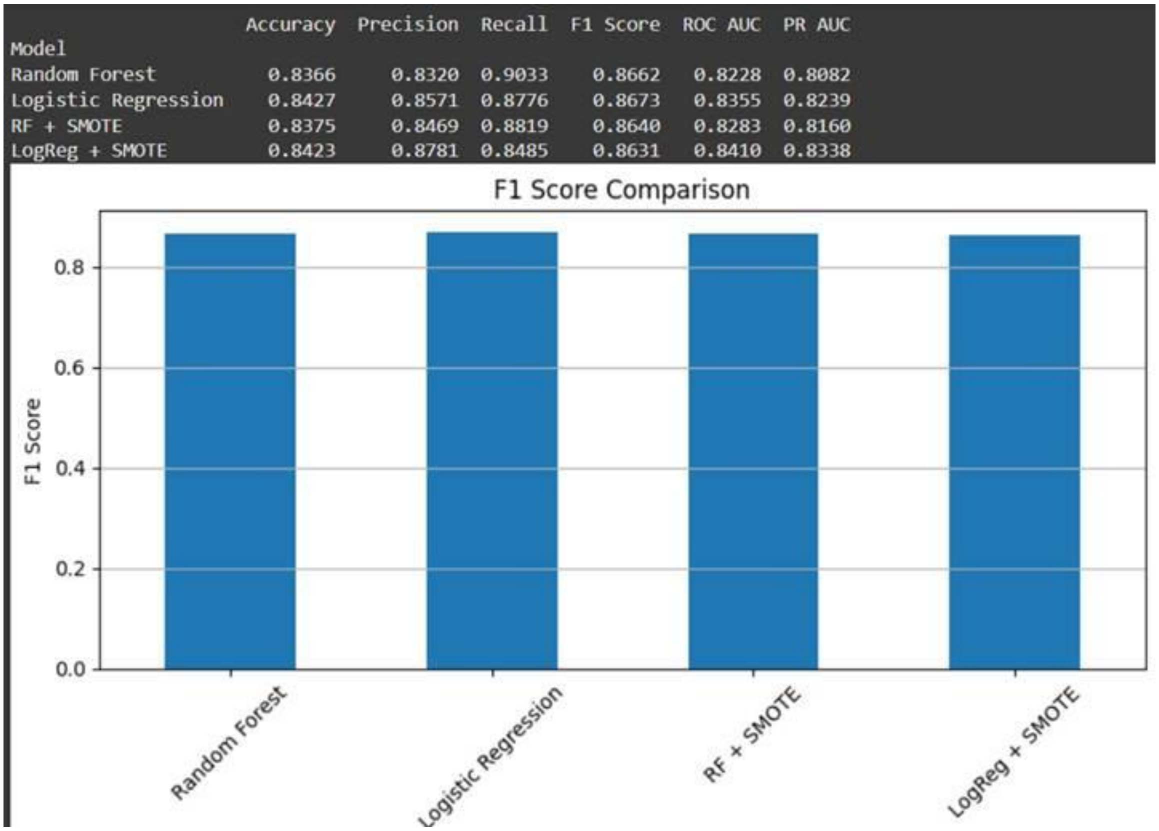
5. Technical Solution

We implemented and compared the following models:

1. Logistic Regression (baseline)
2. Logistic Regression + SMOTE

3. Random Forest
4. Random Forest + SMOTE

SMOTE was used to oversample the minority class to address class imbalance. GridSearchCV with 5-fold Stratified Cross Validation was used for hyperparameter tuning. Logistic Regression (without SMOTE) was ultimately selected due to its strong F1 Score and interpretability.



```
                     Accuracy  Precision  Recall  F1 Score  ROC AUC  PR AUC
Model
Random Forest          0.8366     0.8320  0.9033    0.8662   0.8228  0.8082
Logistic Regression    0.8427     0.8571  0.8776    0.8673   0.8355  0.8239
RF + SMOTE             0.8375     0.8469  0.8819    0.8640   0.8283  0.8160
LogReg + SMOTE         0.8423     0.8781  0.8485    0.8631   0.8410  0.8338
```

Fig 8 Logistic Regression achieved the highest F1 score at 86.7%, making it the preferred model.

Of all the models tried, Logistic Regression without SMOTE performed with the highest F1 score (0.8673), indicating an improved balance between precision (0.8571) and recall (0.8776). While Random Forest was as good at recall (0.9033), it trailed by a margin in precision (0.83320), indicating a higher rate of false-positives. SMOTE, when applied to both Logistic Regression and Random Forest, resulted in improved class balance but added minimal performance dips. For instance, Random Forest with SMOTE suffered a minor F1 Score reduction (0.8640), and Logistic Regression with SMOTE also saw a reduction (0.8631) from the baseline Logistic model.

For ROC AUC and PR AUC, Logistic Regression without SMOTE also led the pack with 0.8355 and 0.8239, respectively. These metrics also confirm that the model not just does a great job at distinguishing depressed from non-depressed students, but also possesses satisfactory performance across different thresholds.

With the need for interpretability in sensitive domains like mental health, Logistic Regression is the hands-down winner. It's coefficient-based model gives one direct access to the effect of every feature, which is rather unobtainable with Random Forest. It was thus selected as the final model, compromising high predictive performance with interpretability-crucial in offering reliable, human-centric recommendations in mental health intervention systems.

6. Feature Importance

We used the magnitude of logistic regression coefficients to evaluate feature importance. Top predictors include suicidal thoughts, academic pressure, and financial stress levels.

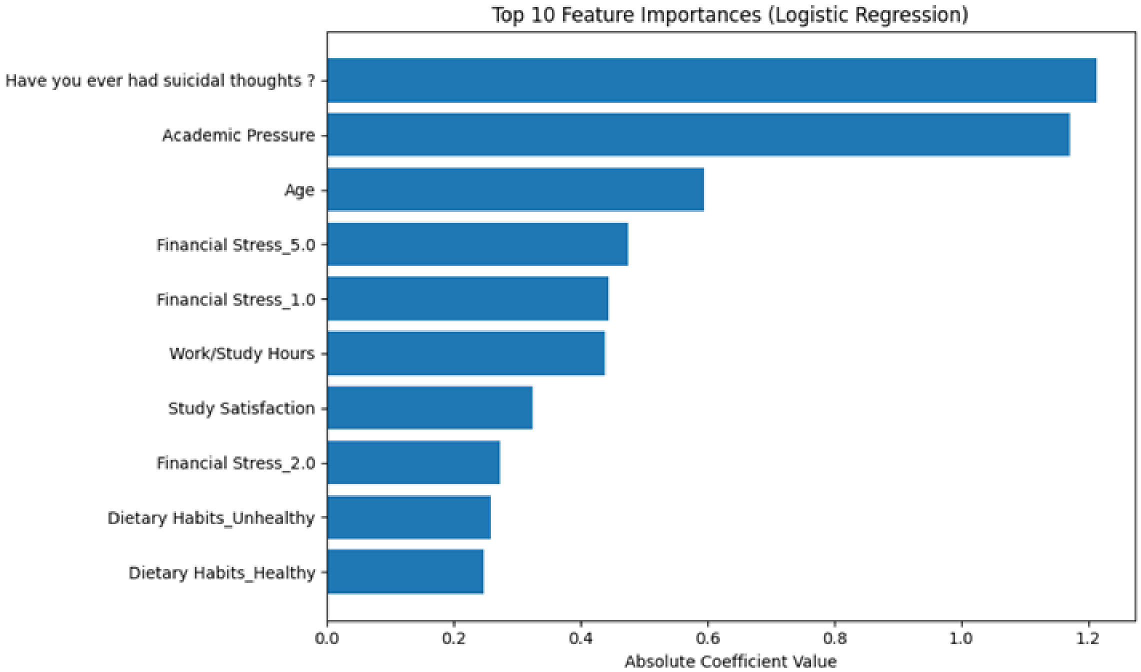
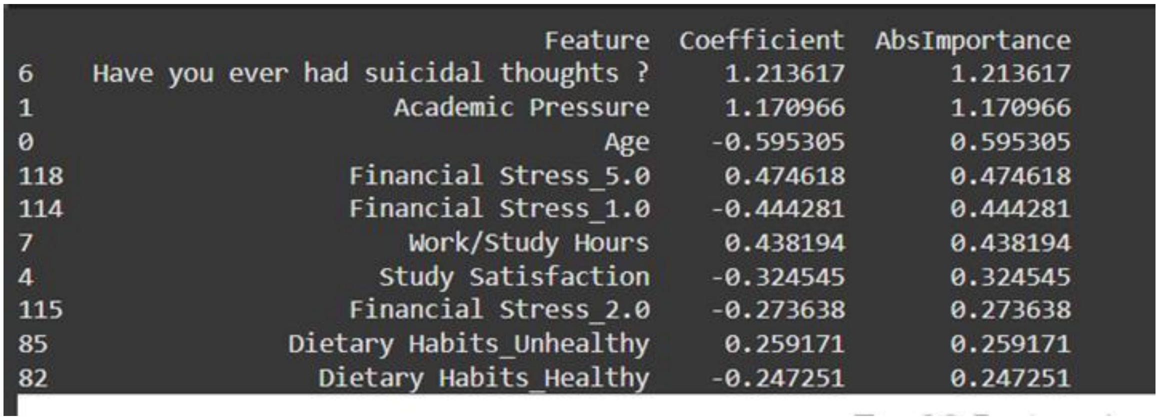


Fig 9: Logistic Regression Feature Importance



```
                                 Feature  Coefficient  AbsImportance
6    Have you ever had suicidal thoughts ?     1.213617       1.213617
1                        Academic Pressure     1.170966       1.170966
0                                      Age    -0.595305       0.595305
118                   Financial Stress_5.0     0.474618       0.474618
114                   Financial Stress_1.0    -0.444281       0.444281
7                         Work/Study Hours     0.438194       0.438194
4                       Study Satisfaction    -0.324545       0.324545
115                   Financial Stress_2.0    -0.273638       0.273638
85                Dietary Habits_Unhealthy     0.259171       0.259171
82                  Dietary Habits_Healthy    -0.247251       0.247251
```

Fig 10: Shows absolute coefficient values for top 10 features used in the model.

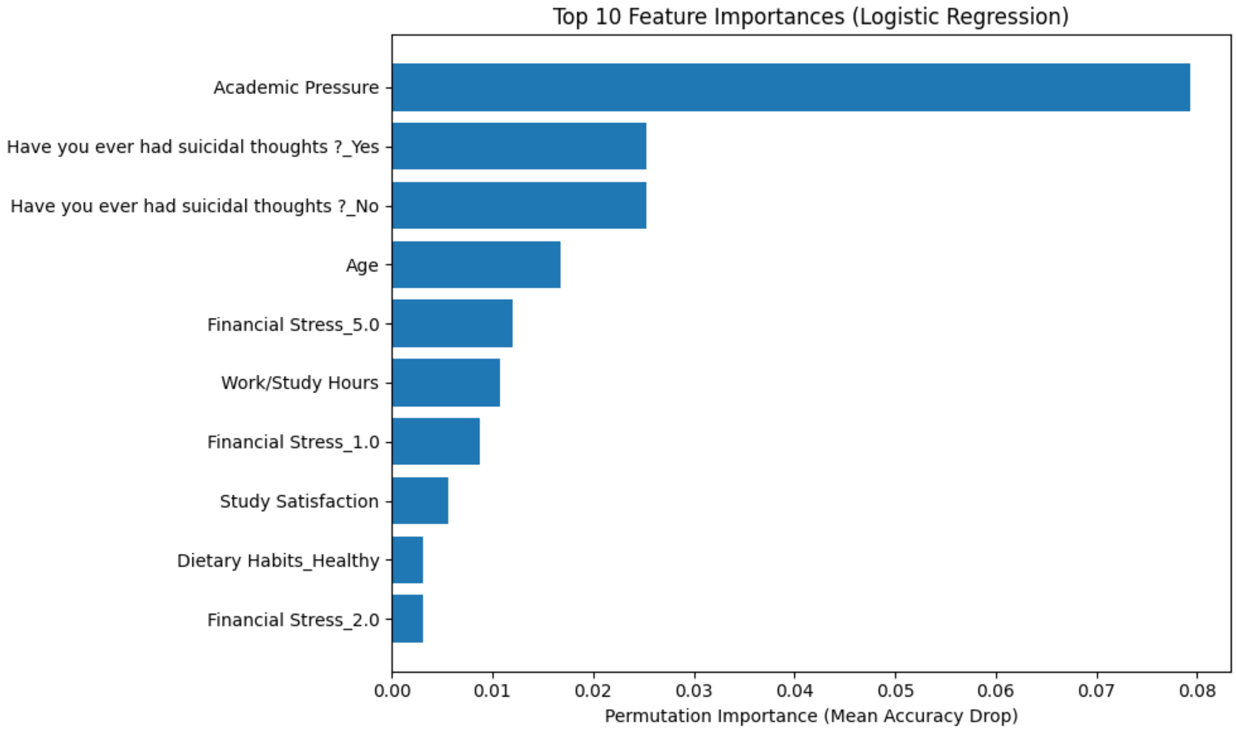
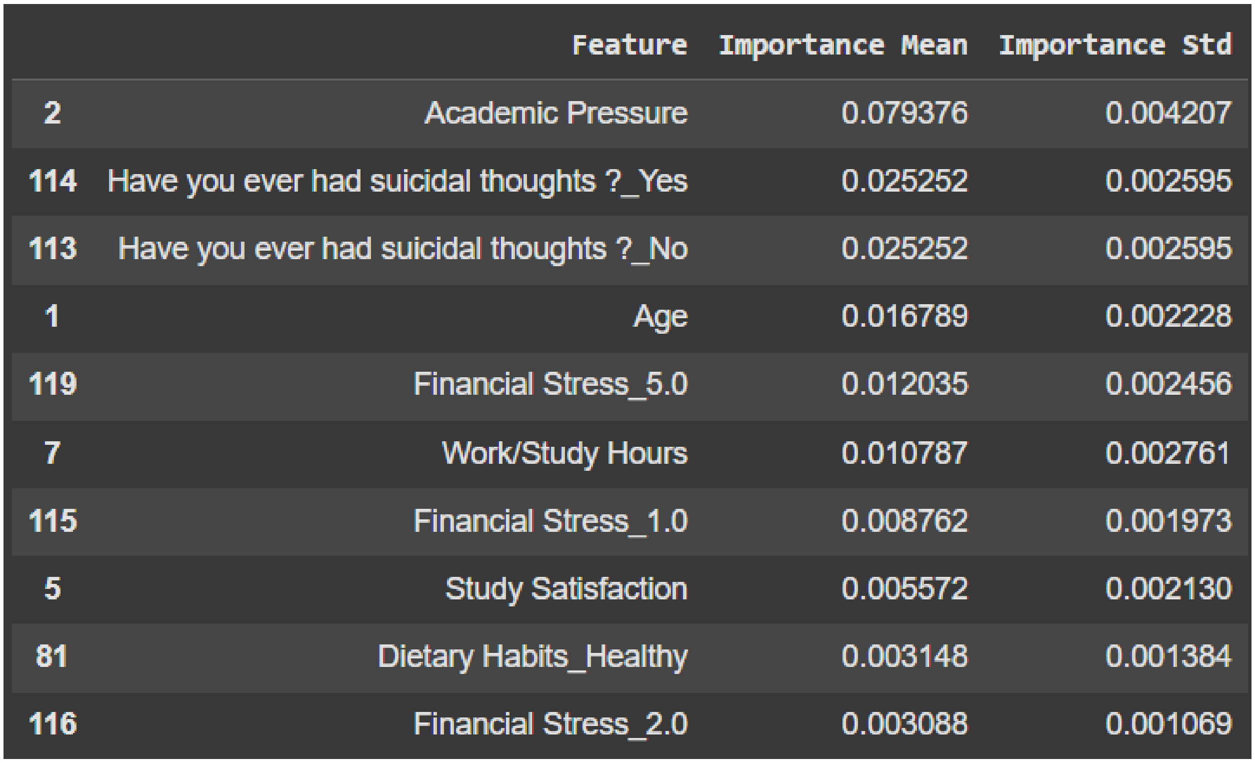


Fig 11: Permutation Importance



	Feature	Importance Mean	Importance Std
2	Academic Pressure	0.079376	0.004207
114	Have you ever had suicidal thoughts ?_Yes	0.025252	0.002595
113	Have you ever had suicidal thoughts ?_No	0.025252	0.002595
1	Age	0.016789	0.002228
119	Financial Stress_5.0	0.012035	0.002456
7	Work/Study Hours	0.010787	0.002761
115	Financial Stress_1.0	0.008762	0.001973
5	Study Satisfaction	0.005572	0.002130
81	Dietary Habits_Healthy	0.003148	0.001384
116	Financial Stress_2.0	0.003088	0.001069

Fig 12: Shows Permutation importance values for top 10 features used in the model.

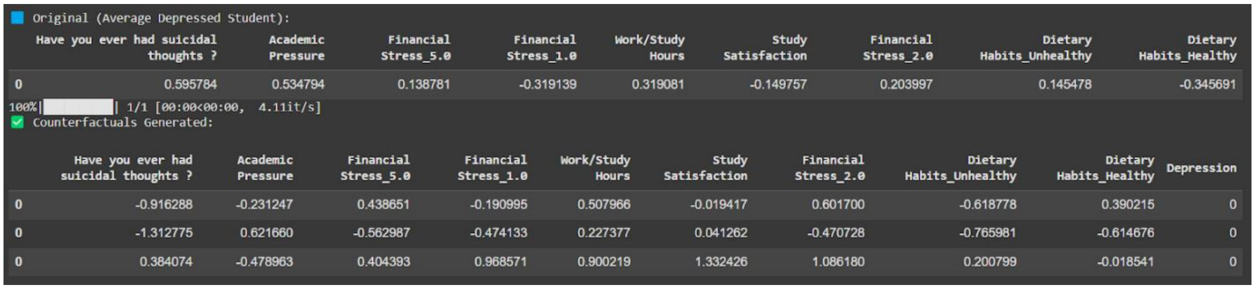
To find the most powerful predictors of depression in our model, we looked at both the permutation importance scores and the absolute logistic regression coefficients. Both were in agreement on Academic Pressure as the most powerful feature, indicating that students under lots of academic pressure are significantly more likely to be depressed. Likewise, responses to suicidal thoughts — and particularly answering "Yes" — were also highly predictive, revealing the strong connection between mental health impairment and depression.

The other significant features were Age, where younger participants were at higher risk of depression, and levels of Financial Stress, which contributed notably to the model's predictions. While coefficients showed each feature's direction and magnitude of its impact, permutation importance better represented each feature's true contribution to model performance in that it accounted for feature interaction and redundancy.

Together, these results provide strong evidence that academic and emotional stressors — specifically, academic pressure and suicidal ideation — are central to the identification of individuals at risk for depression.

7. Counterfactual Generation

We used the DiCE library to generate counterfactuals for 50 students predicted to be depressed. Only 9 editable features were allowed to change. The goal was to find the minimal set of changes needed to flip the prediction to 'not depressed'. Counterfactuals were also generated for the average student profile.



🟦 Original (Average Depressed Student):

	Have you ever had suicidal thoughts ?	Academic Pressure	Financial Stress_5.0	Financial Stress_1.0	Work/Study Hours	Study Satisfaction	Financial Stress_2.0	Dietary Habits_Unhealthy	Dietary Habits_Healthy
0	0.595784	0.534794	0.138781	-0.319139	0.319081	-0.149757	0.203997	0.145478	-0.345691

```
100%|██████████| 1/1 [00:00<00:00,  4.11it/s]
```

✅ Counterfactuals Generated:

	Have you ever had suicidal thoughts ?	Academic Pressure	Financial Stress_5.0	Financial Stress_1.0	Work/Study Hours	Study Satisfaction	Financial Stress_2.0	Dietary Habits_Unhealthy	Dietary Habits_Healthy	Depression
0	-0.916288	-0.231247	0.438651	-0.190995	0.507966	-0.019417	0.601700	-0.618778	0.390215	0
0	-1.312775	0.621660	-0.562987	-0.474133	0.227377	0.041262	-0.470728	-0.765981	-0.614676	0
0	0.384074	-0.478963	0.404393	0.968571	0.900219	1.332426	1.086180	0.200799	-0.018541	0

Fig 13: Counterfactual for Average Student; Three counterfactual examples show changes that lead to a prediction of 'not depressed'.

For instance, we focused on the standard profile of a depressed student with high scores for academic pressure, financial stress, and suicidal ideation, and unhealthy eating patterns and low study satisfaction. We used the DiCE library to generate counterfactuals by modifying only nine editable features in order to ensure realism and feasibility.

The three example counterfactuals in the figure illustrate alternative but plausible paths to changing the model's prediction from depressed to not depressed. Common changes include reducing the amount of academic pressure, financial hardship, and work/study hours, and increasing study satisfaction and eating behavior. Curiously, the experience of having suicidal thoughts had to be changed significantly for the prediction to change – underscoring it's salience in mental health assessment.

These counterfactuals specify how small, targeted interventions can reduce depression risk. For instance, academic advising to reduce stress, financial aid services, mental health services to treat suicidal thoughts, and wellness programs that promote healthier lifestyle options could be recommended. The ability to generate such customized, actionable recommendations transforms the model from a diagnostic tool to a prescriptive support system for student well-being.

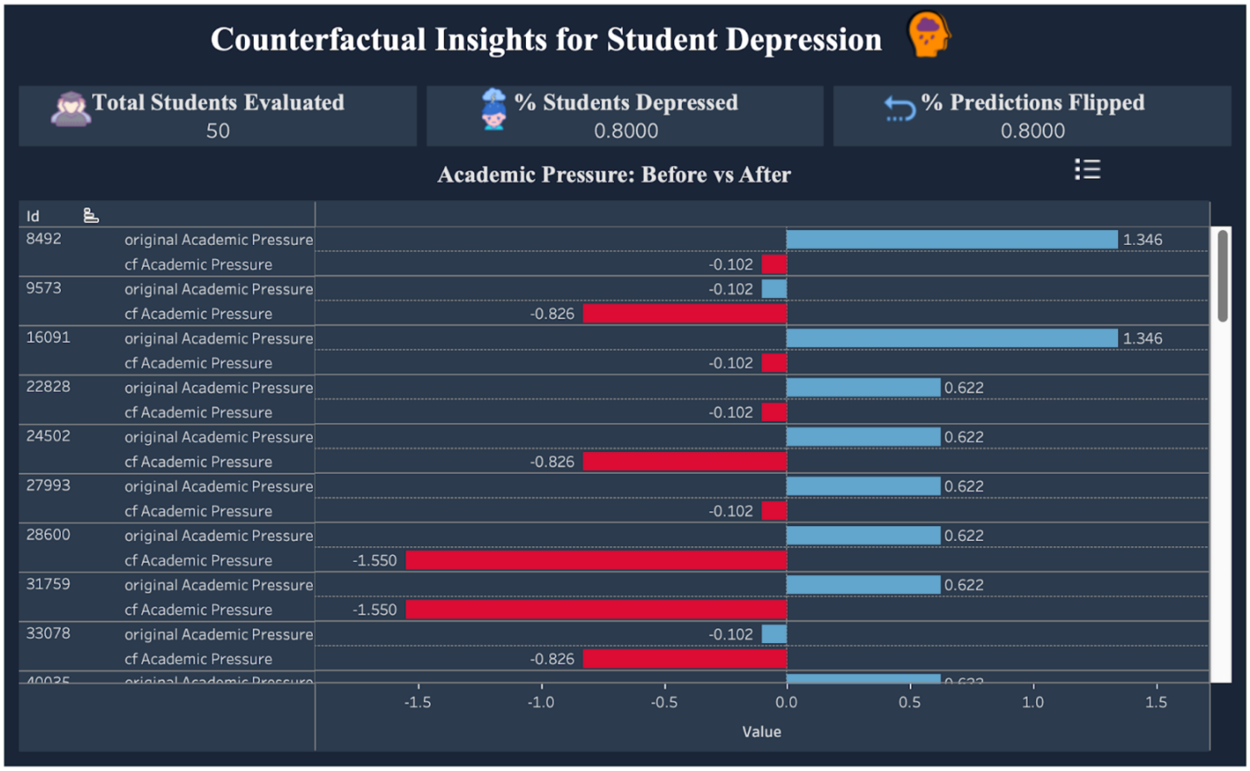
8. Results

Final Model Performance (Logistic Regression)

Metric	**Value**
Accuracy	84.3%
Precision	87.8%
Recall	84.5%
F1 Score	86.7%
ROC AUC	83.6%
PR AUC	82.4%

Most frequently modified features in counterfactuals: Academic Pressure, Financial Stress, Suicidal Thoughts.

Dashboard Screenshots:



Counterfactual Insights for Student Depression

Total Students Evaluated
50

% Students Depressed
0.8000

% Predictions Flipped
0.8000

Academic Pressure: Before vs After

Fig 14: How academic pressure impacts predictions: Before and After Intervention

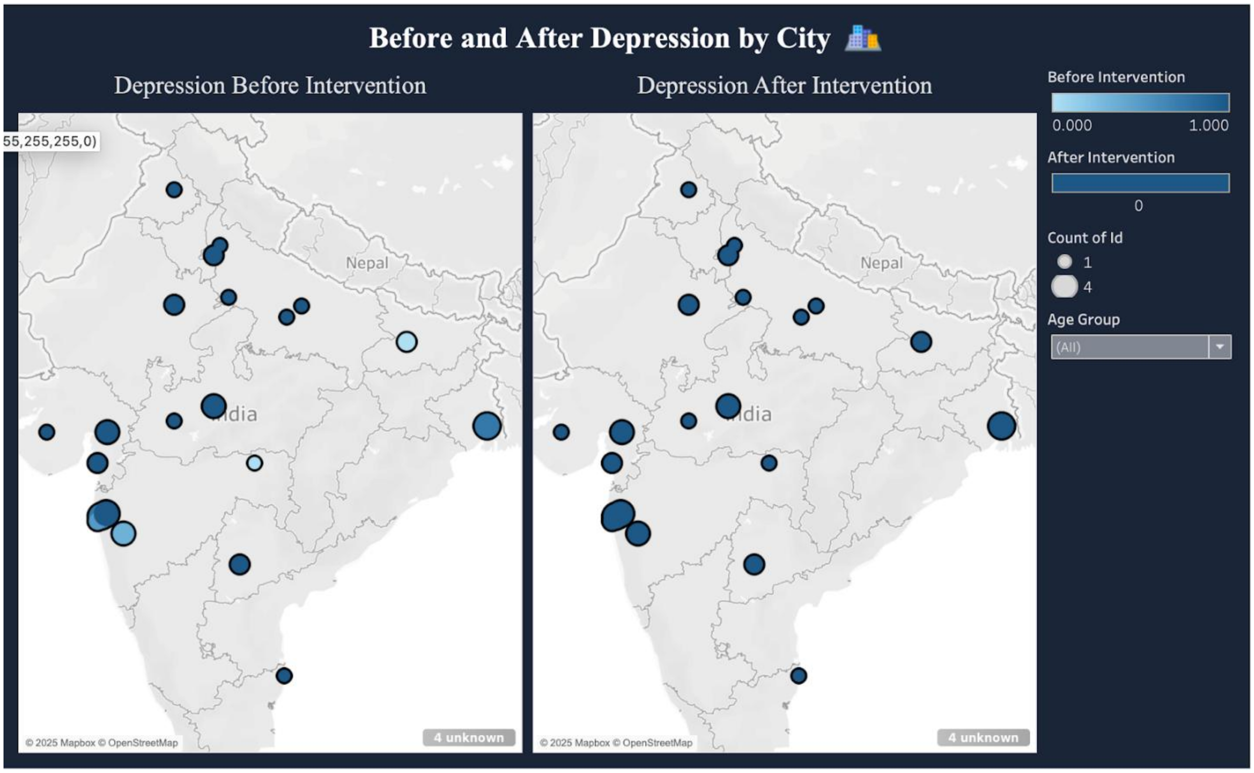
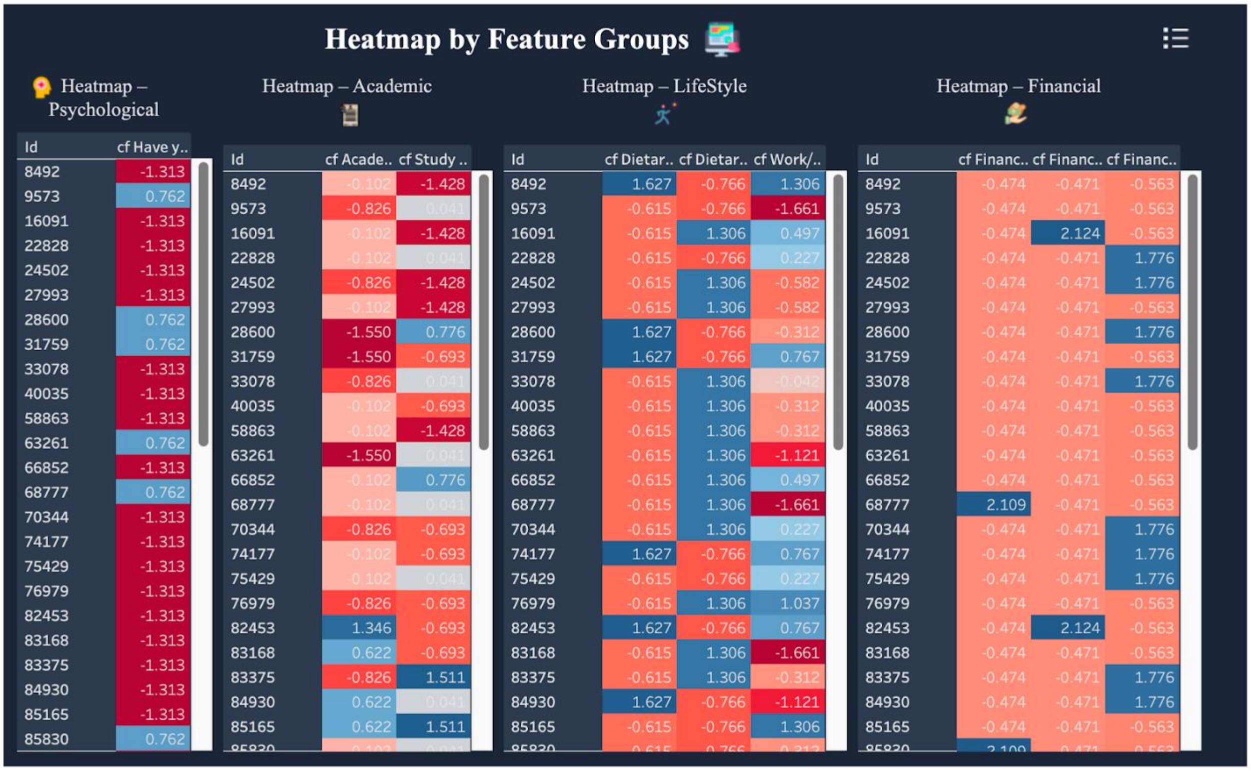

Before and After Depression by City
Depression Before Intervention
Depression After Intervention
55,255,255,0)
Nepal
India
© 2025 Mapbox © OpenStreetMap
4 unknown
Before Intervention
0.000
1.000
After Intervention
0
Count of Id
1
4
Age Group
(All)

Fig 15: Geographic impact of depression: A city-wise before and after comparison.



Heatmap – Psychological

Id	cf Have y..
8492	-1.313
9573	0.762
16091	-1.313
22828	-1.313
24502	-1.313
27993	-1.313
28600	0.762
31759	0.762
33078	-1.313
40035	-1.313
58863	-1.313
63261	0.762
66852	-1.313
68777	0.762
70344	-1.313
74177	-1.313
75429	-1.313
76979	-1.313
82453	-1.313
83168	-1.313
83375	-1.313
84930	-1.313
85165	-1.313
85830	0.762

Heatmap – Academic

Id	cf Acade..	cf Study ..
8492	-0.102	-1.428
9573	-0.826	0.041
16091	-0.102	-1.428
22828	-0.102	0.041
24502	-0.826	-1.428
27993	-0.102	-1.428
28600	-1.550	0.776
31759	-1.550	-0.693
33078	-0.826	0.041
40035	-0.102	-0.693
58863	-0.102	-1.428
63261	-1.550	0.041
66852	-0.102	0.776
68777	-0.102	0.041
70344	-0.826	-0.693
74177	-0.102	-0.693
75429	-0.102	0.041
76979	-0.826	-0.693
82453	1.346	-0.693
83168	0.622	-0.693
83375	-0.826	1.511
84930	0.622	0.041
85165	0.622	1.511

Heatmap – LifeStyle

Id	cf Dietar..	cf Dietar..	cf Work/..
8492	1.627	-0.766	1.306
9573	-0.615	-0.766	-1.661
16091	-0.615	1.306	0.497
22828	-0.615	-0.766	0.227
24502	-0.615	1.306	-0.582
27993	-0.615	1.306	-0.582
28600	1.627	-0.766	-0.312
31759	1.627	-0.766	0.767
33078	-0.615	1.306	-0.042
40035	-0.615	1.306	-0.312
58863	-0.615	1.306	-0.312
63261	-0.615	1.306	-1.121
66852	-0.615	1.306	0.497
68777	-0.615	1.306	-1.661
70344	-0.615	1.306	0.227
74177	1.627	-0.766	0.767
75429	-0.615	-0.766	0.227
76979	-0.615	1.306	1.037
82453	1.627	-0.766	0.767
83168	-0.615	1.306	-1.661
83375	-0.615	1.306	-0.312
84930	1.627	-0.766	-1.121
85165	-0.615	-0.766	1.306

Heatmap – Financial

Id	cf Financ..	cf Financ..	cf Financ..
8492	-0.474	-0.471	-0.563
9573	-0.474	-0.471	-0.563
16091	-0.474	2.124	-0.563
22828	-0.474	-0.471	1.776
24502	-0.474	-0.471	1.776
27993	-0.474	-0.471	-0.563
28600	-0.474	-0.471	1.776
31759	-0.474	-0.471	-0.563
33078	-0.474	-0.471	1.776
40035	-0.474	-0.471	-0.563
58863	-0.474	-0.471	-0.563
63261	-0.474	-0.471	-0.563
66852	-0.474	-0.471	-0.563
68777	2.109	-0.471	-0.563
70344	-0.474	-0.471	1.776
74177	-0.474	-0.471	1.776
75429	-0.474	-0.471	1.776
76979	-0.474	-0.471	-0.563
82453	-0.474	2.124	-0.563
83168	-0.474	-0.471	-0.563
83375	-0.474	-0.471	1.776
84930	-0.474	-0.471	1.776
85165	-0.474	-0.471	-0.563

Fig 16: Grouped counterfactuals: Patterns across psychological, academic, lifestyle, and financial domains

Dashboard Link:

https://public.tableau.com/views/DV_Project_Counterfactual_Insights_for_Student_Depression/Story1?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

9. Conclusion

This project integrates interpretable machine learning with counterfactual explanations to improve student mental health prediction. Logistic Regression was selected for its transparency, and DiCE was used to suggest actionable changes. The approach is practical and scalable and can be further extended using clinical feedback or real-time dashboards. Limitations include static data and lack of longitudinal tracking.

To generate these results, we used a tool called **DiCE**, short for Diverse Counterfactual Explanations. It allowed us to create multiple realistic alternatives for a student labeled as depressed. These counterfactuals highlight minimal, personalized changes — like improved diet, better study satisfaction, or reduced financial stress — that could shift a student's mental health prediction in a positive direction. The goal is to turn black-box AI into something interpretable, useful, and human-centered.

10. Summary

This project presents a realistic approach to predicting student mental health based on interpretable machine learning and counterfactual analysis. By using Logistic Regression and DiCE, we demonstrated that not only can we accurately predict students who are at risk of depression but also present realistic and personalized intervention recommendations. The results highlight academic pressure, suicidal thoughts, and financial struggles as key drivers toward the risk of depression. With an F1 score of 86.7%, our model is both performant and interpretable. Visual dashboards and counterfactuals allow stakeholders to gain insightful and actionable information that is easy to interpret. This work provides a foundation for the use of explainable, student-centered AI tools in education and guidance environments.

11. Future Work

Future directions for this project involve:

1. Longitudinal Data Integration: Incorporating time-series or multi-semester data to monitor changes in student mental health across time.
2. Feature Augmentation: Incorporating additional contextual features such as social support networks, extracurricular activities, and clinical diagnosis history for better prediction.
3. Real-Time Monitoring System: Developing a live dashboard integrated with institutional systems for real-time depression risk alerts and monitoring interventions.
4. Feedback-Loop with Counselors: Collaboration with mental health professionals to verify counterfactual recommendations and fine-tune them for clinical usability.
5. Bias and Fairness Analysis: Evaluating model bias among various demographic groups (e.g., gender, city, degree level) and applying fairness-aware algorithms if inequalities exist.
6. Alternative Explanation Methods: Contrasting DiCE with SHAP, LIME, or anchor-based explanations to evaluate robustness and user trust in recommendations.
7. Deployment in Educational Settings: Pilot testing the model and dashboard within actual schools or universities to try adoption, usability, and student outcomes.

12. References

1. Molnar, C. (2022). Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/
2. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology.
3. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys (CSUR), 51(5), 1-42.
4. Looveren, A. V., & Klaise, J. (2021). Interpretable Counterfactual Explanations Guided by Prototypes. arXiv preprint arXiv:2106.06850.
5. DiCE Library: https://github.com/interpretml/DiCE
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
7. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.