
Toward Early Detection: A Model for Student Depression Risk

Sahil Sangani, Jonathan Tang, Ruosi Hu, Vera Feng



Introduction



Research Goal

Primary Research Goal:

- Effectively predict the presence of Depression of students for new datasets

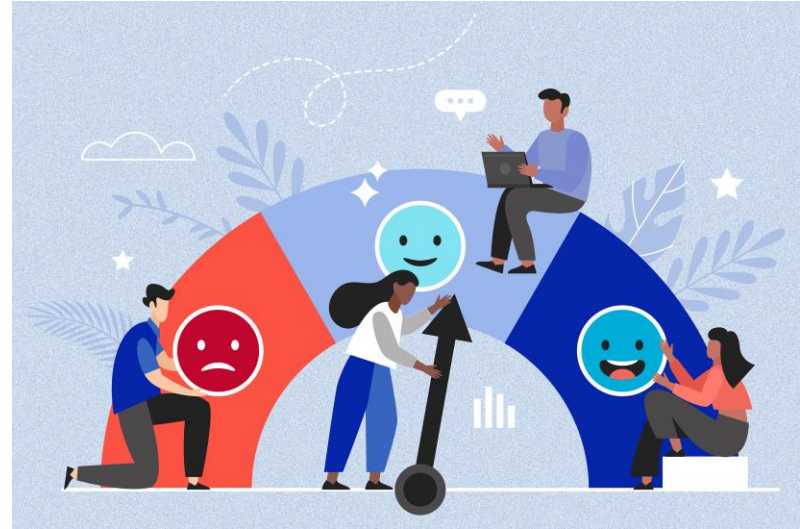
Secondary research Goal:

- Yield reliable interpretative insights about the nature of the relationship between the variables in the dataset

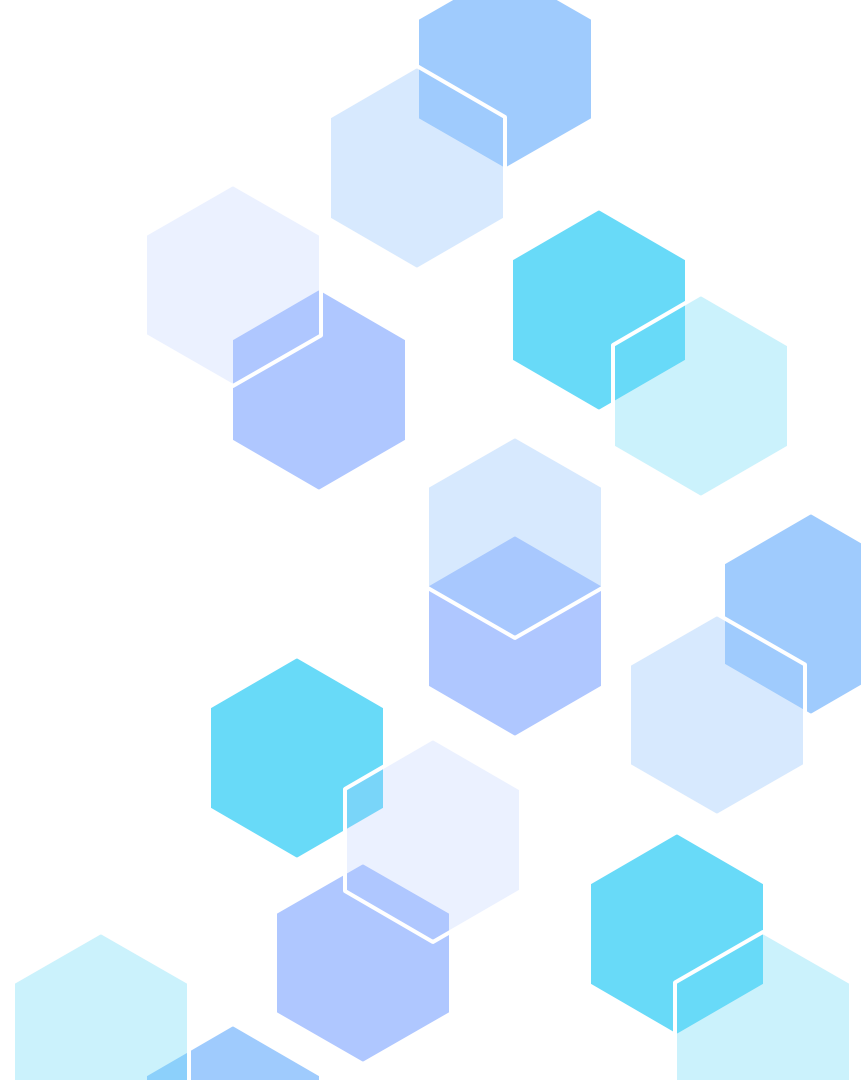


Motivation

- Predict depression from simple survey
- Fast, low-cost mental health screening
- Identify at-risk students early



Dataset Overview



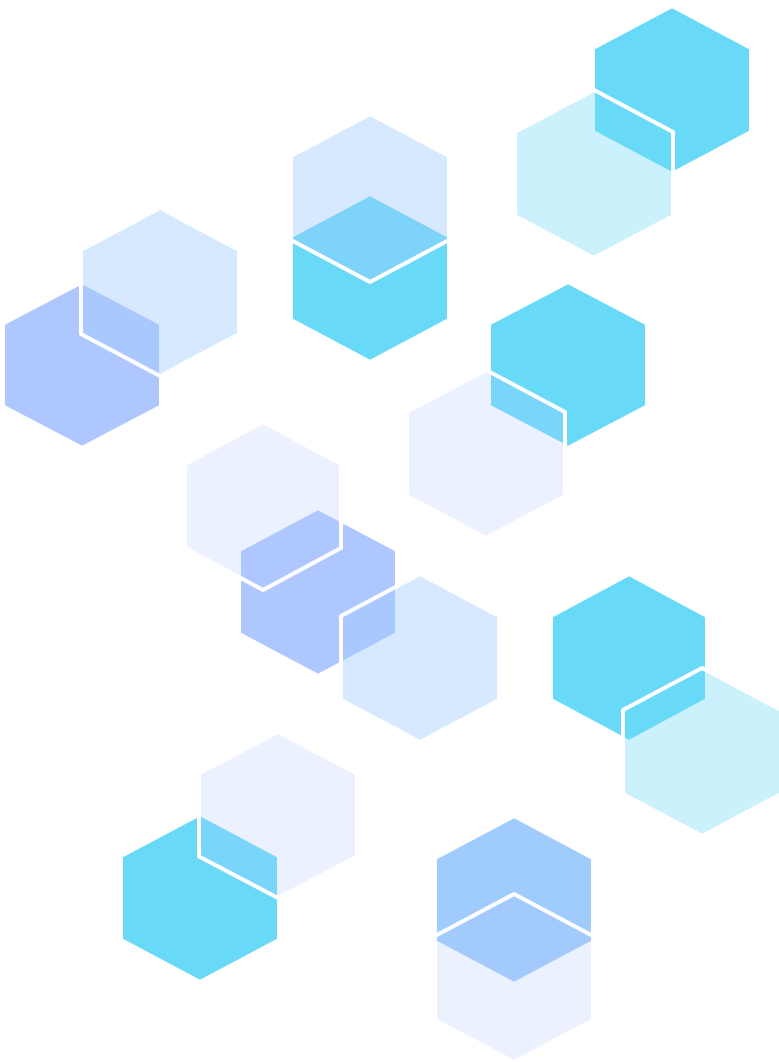
Dataset Source

- **Kaggle link:** <https://www.kaggle.com/datasets/ikynahidwin/depression-student-dataset>
- **Download date:** 04/22/2025
- **Result source:** A Survey

Dataset Display

	Gender	Age	Academic Pressure	Study Satisfaction	Sleep Duration	Dietary Habits	Study Hours	Financial Stress	Depression
0	Male	28	2.0	4.0	7-8 hours	Moderate	9	2	No
1	Male	28	4.0	5.0	5-6 hours	Healthy	7	1	No
2	Male	25	1.0	3.0	5-6 hours	Unhealthy	10	4	Yes
3	Male	23	1.0	4.0	More than 8 hours	Unhealthy	7	2	No
4	Female	31	1.0	5.0	More than 8 hours	Healthy	4	2	No

- **Dataset Attribute:** 502 rows
- **Explanatory Variables:** First 8 columns (gender, age, academic pressure, etc.)
- Self-perception scores are **discrete**.
- **Response Variable:** Last 1 column - 'Depression'



Main features

- Target Array: column of Depression level
- Feature Matrix: All columns except Depression level
- Will use standard scaler to ensure similar STD



Data Cleaning

- No missing values
- No outliers
- No change to the dataset

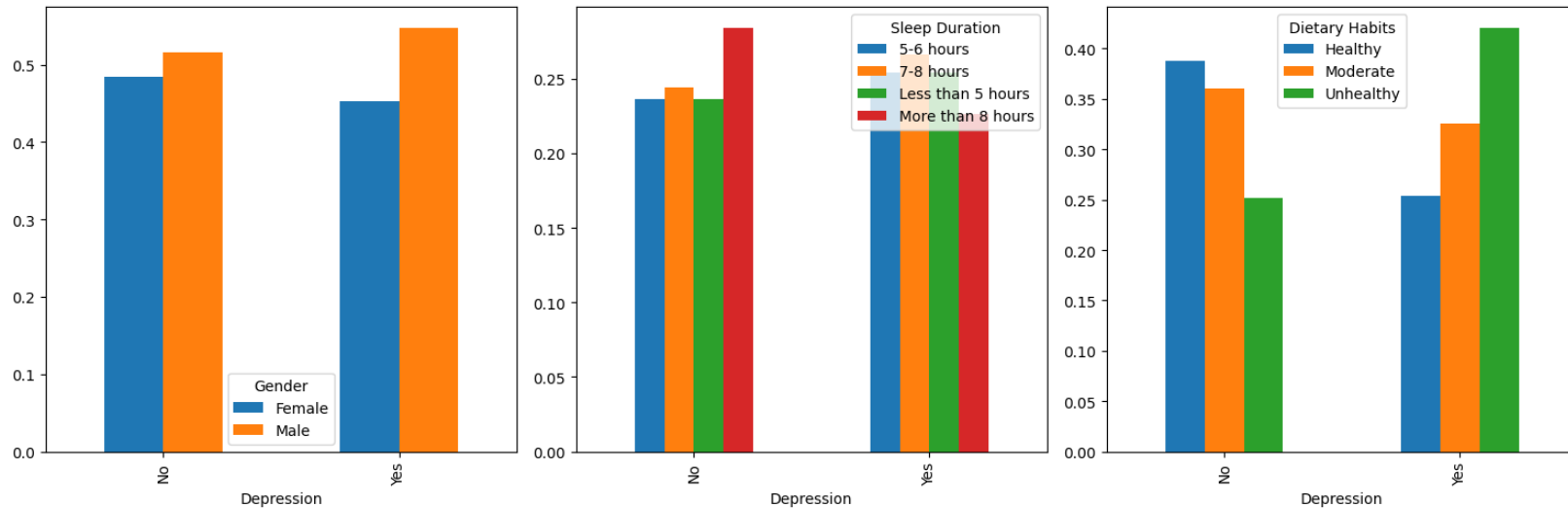


Preliminary Analysis

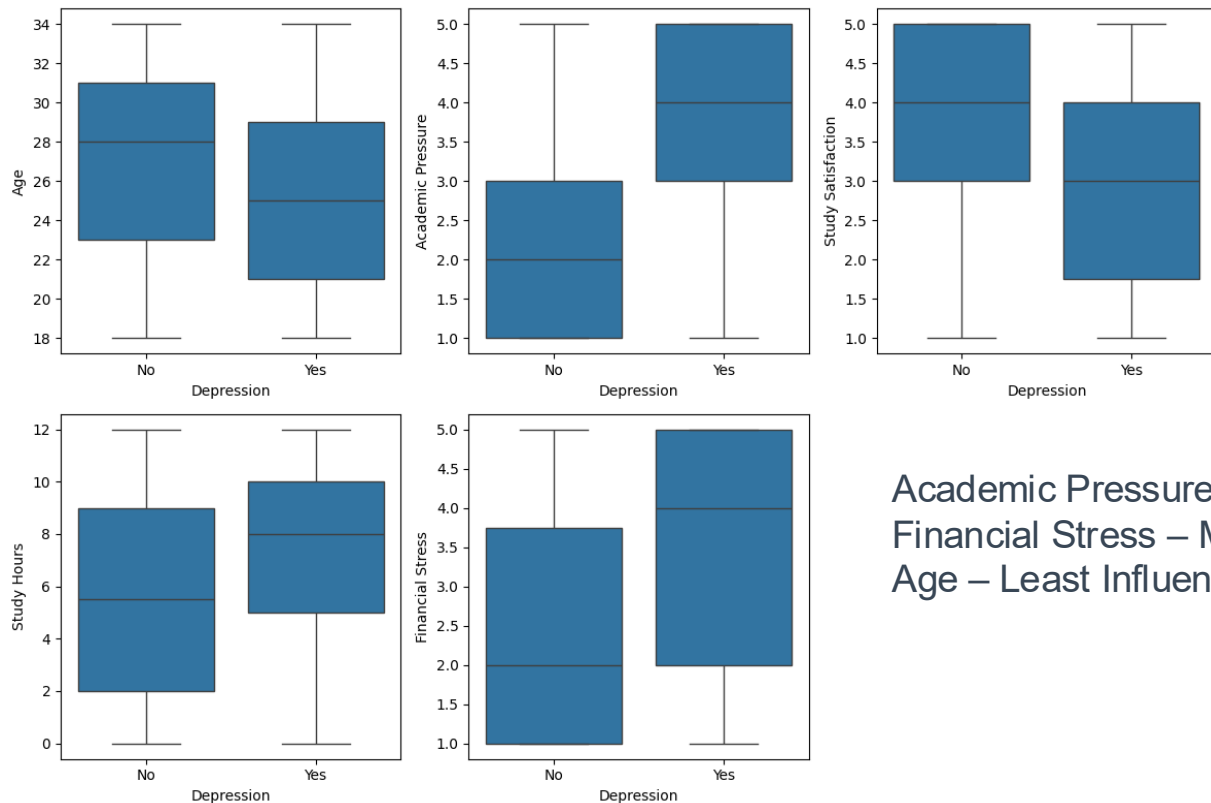
- Identify relationships of variables
- Choose a regularization technique

Depression vs Categorical Variables

Dietary Habits – Most Influential
Gender – Least Influential

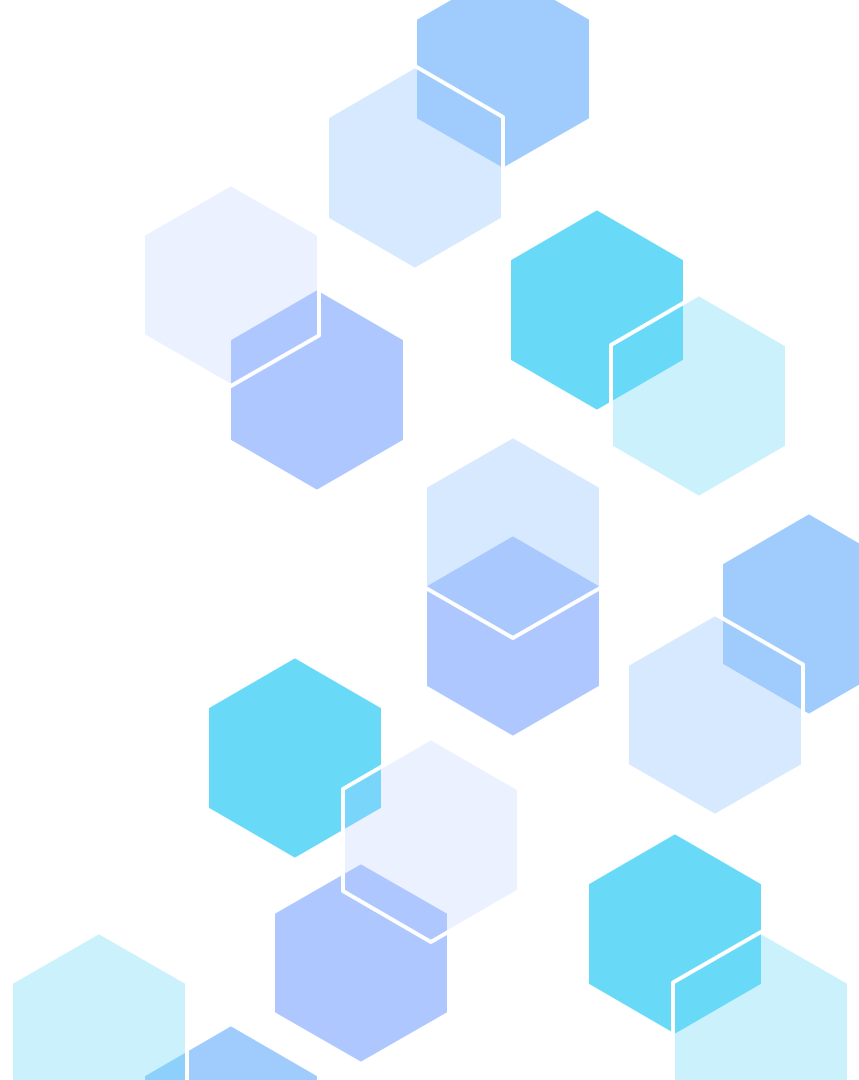



Depression vs Numerical Variables



Academic Pressure and
Financial Stress – Most Influential
Age – Least Influential

Model Evaluation





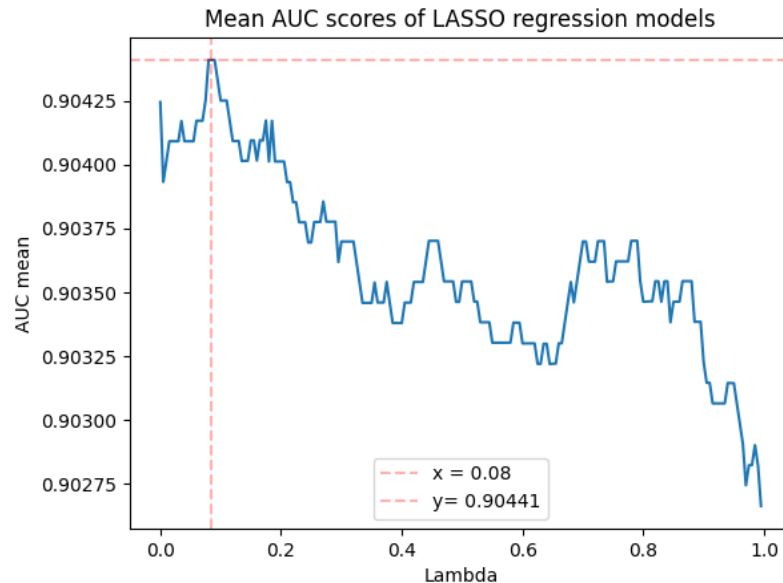
Model Selection

- Used Lasso (L1) Logistic Regression
- Simplifies model by zeroing unimportant features
- Improves interpretability & reduces survey length

Comparison of model performance

Our best model AUC mean is 0.90441

Full model AUC mean is 0.90425





Best Model Analysis

- The lambda value 0.08 maximized AUC mean
- Lasso selection & Logistic regression model

Lasso Selection

- No overfitting, no variables zeroed out
- Dietary Habits is the most important variable

	lasso_mod_08
Age	-0.873425
Academic Pressure	1.589962
Study Satisfaction	-1.055172
Study Hours	0.776754
Financial Stress	1.004374
Gender_Male	0.013350
Sleep Duration_7-8 hours	0.043153
Sleep Duration_Less than 5 hours	0.359744
Sleep Duration_More than 8 hours	-0.393740
Dietary Habits_Moderate	0.773921
Dietary Habits_Unhealthy	1.867301

Model Equation

Reference Levels:

Gender – Female

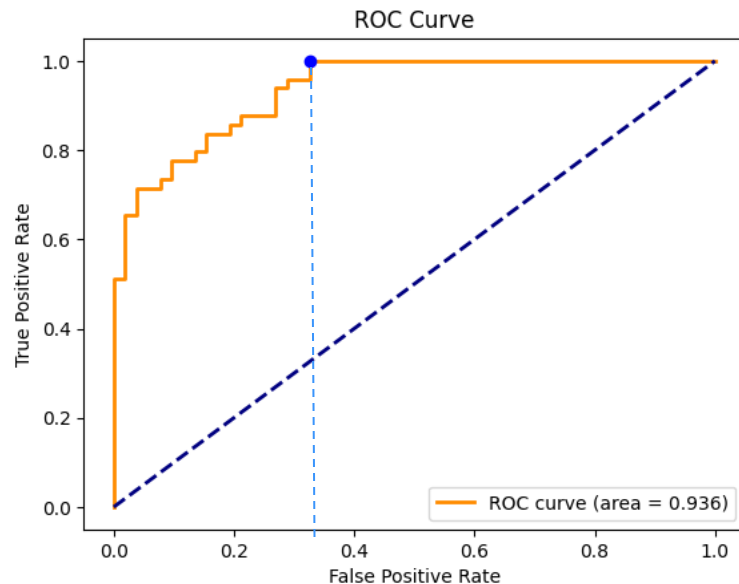
Sleep Duration – 5-6 Hours

Dietary Habits - Healthy

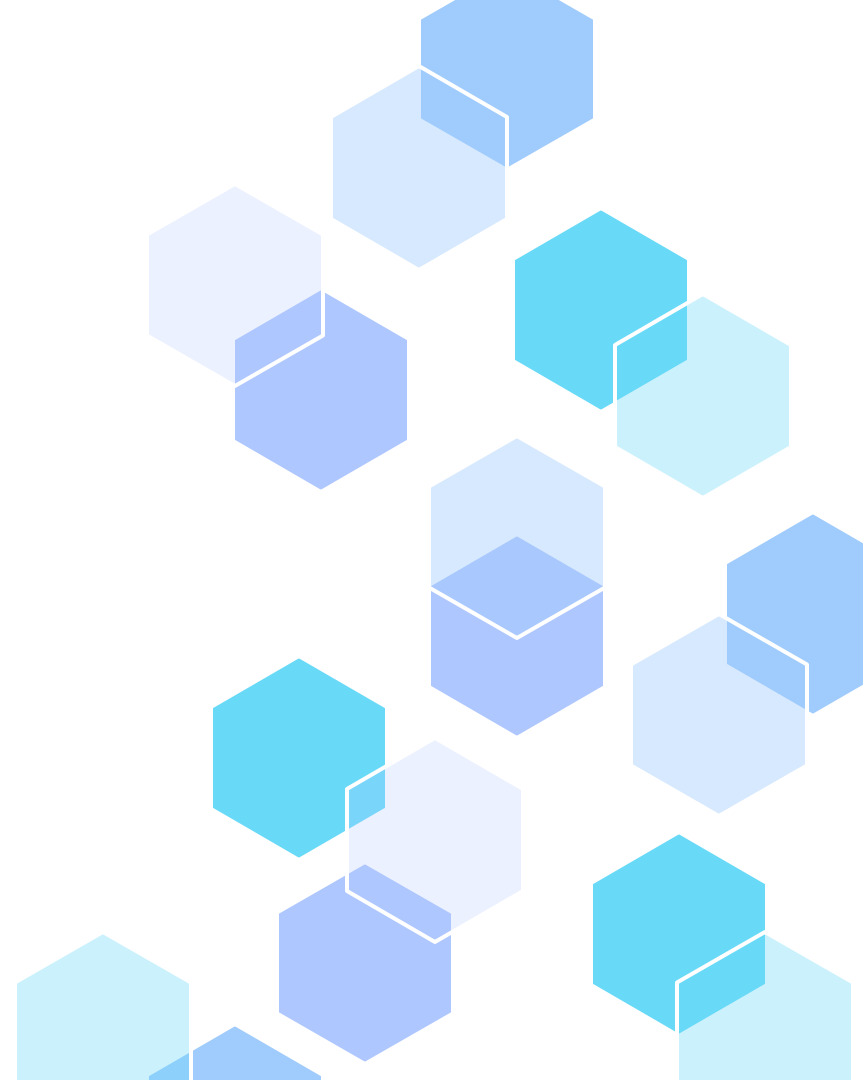
$$\hat{p} = \frac{1}{1 + \exp \left(\begin{array}{l} - 0.8045 \\ - 0.8734 \times \text{Age} \\ + 1.5899 \times \text{Academic Pressure} \\ - 1.0552 \times \text{Study Satisfaction} \\ + 0.7767 \times \text{Study Hours} \\ + 1.0044 \times \text{Financial Stress} \\ + 0.0133 \times \text{Gender}_{\text{Male}} \\ + 0.0434 \times \text{Sleep Duration}_{7-8 \text{ hours}} \\ + 0.3600 \times \text{Sleep Duration}_{\text{Less than 5 hours}} \\ - 0.3935 \times \text{Sleep Duration}_{\text{More than 8 hours}} \\ + 0.7738 \times \text{Dietary Habits}_{\text{Moderate}} \\ + 1.8670 \times \text{Dietary Habits}_{\text{Unhealthy}} \end{array} \right)}$$

ROC Curve Analysis

- An AUC of 0.936 indicates excellent model performance
- Maximizing TPR (rate of correctly identifying depressed students)
- Threshold = 0.2316 yields TPR of 1.0 and FPR of 0.33 (Moderate)

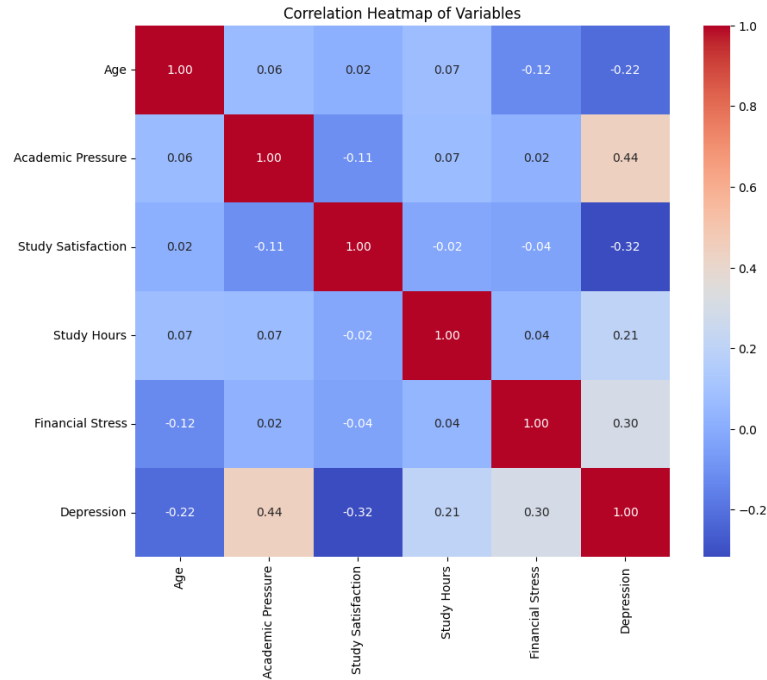


Additional Analysis and Insights



No Multicollinearity

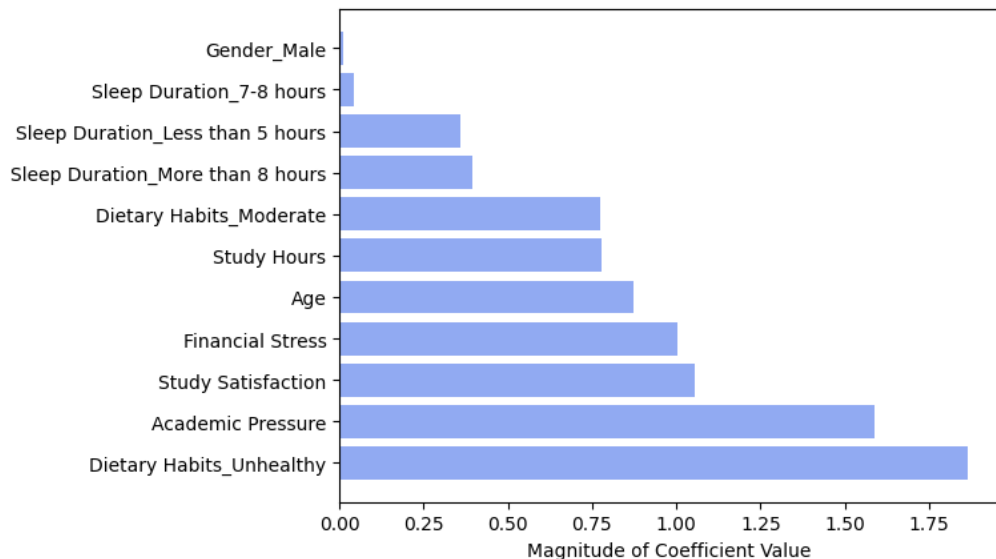
- No strong variable correlations
- Coefficients remain interpretable



Predictive Power of Explanatory Variables

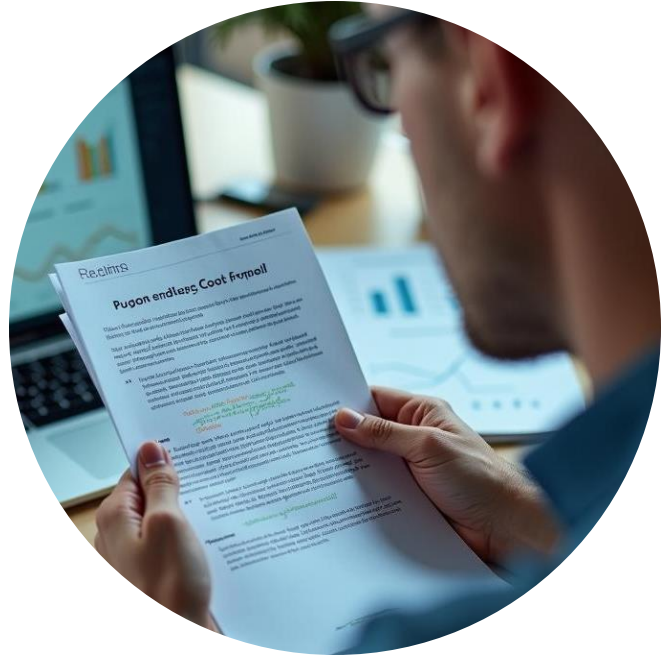
Coefficient magnitude reflects each variable's predictive power

- Standardized numerical variables
- No multicollinearity



Conclusions

- We recommend our best model because our final AUC score is 0.936
- Effective in finding depressed students
- Alternative method: Ridge regression and Elastic Net models
- Limitations and future work: May lack objectivity based on the survey



Thank you!

Reference

American Academy of Pediatrics. (2022, November 15). Supporting students with depression in school. Home. https://www.aap.org/en/patient-care/school-health/mental-health-in-schools/supporting-students-with-depression-in-school/?srsltid=AfmBOorhCOW_Vbe3TI7HiNoWkdGVHB6xw6wE6JGoNaSougkQo2ISi97r
Lochrie, A. S. (Ed.). (2022, April). Depression Factsheet (for Schools) (for parents) | nemours kidshealth. KidsHealth. <https://kidshealth.org/en/parents/depression-factsheet.html>

CREDITS: This presentation template was created by Slidesgo, and includes icons by Flaticon, and infographics & images by Freepik