**Student depression study**

# **Dataset description**

Dataset source: <https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset/data>

**Overview**This dataset compiles a wide range of information aimed at understanding, analyzing, and predicting depression levels among students. It is designed for research in psychology, data science, and education, providing insights into factors that contribute to student mental health challenges and aiding in the design of early intervention strategies. The dataset is cross-sectional, capturing data at a single point in time. We chose this dataset due to the significance of mental health in educational settings and the potential to identify factors associated with depression.

For this project, we have used ChatGPT to help revise and edit our R code in order to make sure the models and variables are properly formatted and fitting for analysis, especially due to issues such as our target variable being binary.

**Data Description**

As per the description of the dataset provided by its author, Adil Shamim:

* **Format:** CSV (each row represents an individual student)
* **Features:**
  + **ID:** Unique identifier for each student
  + **Demographics:** Age, Gender, City
  + **Academic Indicators:** CGPA, Academic Pressure, Study Satisfaction
  + **Lifestyle & Wellbeing:** Sleep Duration, Dietary Habits, Work Pressure, Job Satisfaction, Work/Study Hours
  + **Additional Factors:** Profession, Degree, Financial Stress, Family History of Mental Illness, and whether the student has ever had suicidal thoughts
* **Target Variable:**
  + **Depression\_Status:** A binary indicator (0/1 or Yes/No) that denotes whether a student is experiencing depression

To keep the model simple, we decided to use 3 variables from this set on top of the target variable. Some of the headings for the variables have been changed in the dataset to make working with R easier.

# **Analysis explanation**

**Aim of the Analysis**

The main goal of this analysis is to predict depression levels in students based on lifestyle features such as sleep, diet, and the strength of academic pressure. We aim to compare the performance of Lasso and Ridge regression models, evaluate their predictive accuracy using AUC (Area Under the Curve), and assess which variables contribute most to the prediction.

**Methodology**

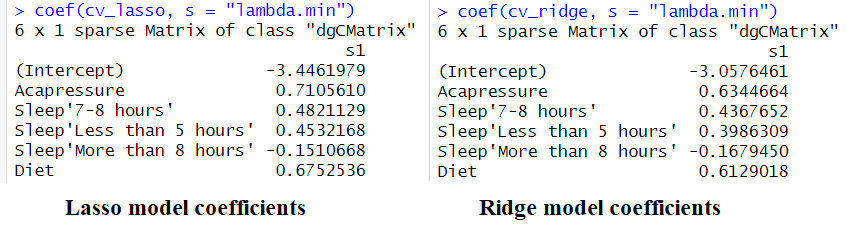
As the dataset was very large, we limited it to 450 randomly chosen responses.

We used the “glmnet” package in R to fit both Lasso (L1) and Ridge (L2) regression models. After removing NA values, we converted categorical variables appropriately and created a model matrix with predictors. The data was split into training (80%) and testing (20%) sets. Cross-validation was applied to select the optimal penalty (lambda) for each model.

# **Interpretation of the results**

Both the ridge and lasso methods yielded similar results – both have an AUC over 0.82, suggesting the data does not have discrimination. On top of that, their accuracy is reasonable at 73%.

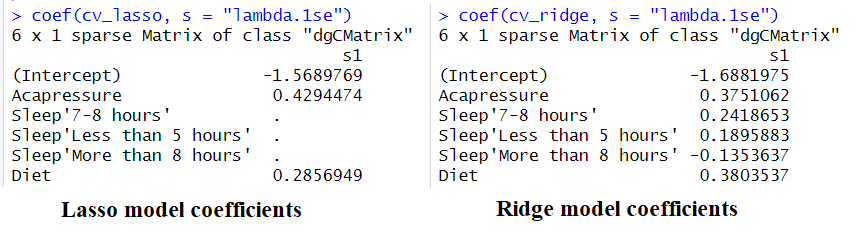
For both models, the results using the **lowest level of lambda** are as follows:



Interpretations of variables:

* Academic pressure (‘acapressure’) is ordered 1-5, with 5 being the worst, meaning the higher the academic pressure, the higher the likelihood of the student having depression.
* Diets are ordered 1-4, with 4 being the worst, meaning the higher the diet, the higher the likelihood of the student having depression.
* Sleep is a dummy variable compared to 5-6 hours of sleep. Meaning, compared to the student getting 5-6 hours of sleep, they are at a bigger risk of having depression when they sleep 7-8 hours or less than 5 hours. However, comparatively, they are less likely to have depression if they sleep more than 8 hours.

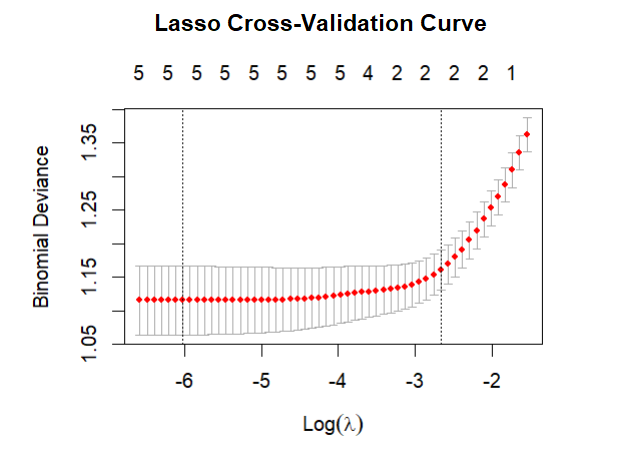
For both models, the results using the stricter interpretation, aka the **more regularized lambda.1se** are as follows:

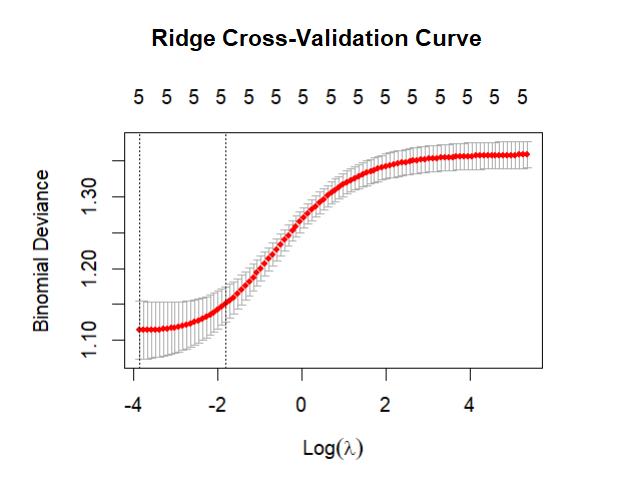


Here, there are slight differences between the Ridge and Lasso model coefficients. As the sleep variables have a very low effect according to the Lasso model, they have been automatically redacted from the results as their coefficient is close to 0.

Interpretations of variables:

* All variables have been kept in the models, but their effect has shrunk.





#### **Lasso Model (Top Plot):**

* The curve remains relatively flat at lower values of log(λ), meaning the model maintains stable performance over a wide range of regularization strengths.
* There is a clear minimum point, beyond which the deviance begins to rise – this indicates that overly large λ values hurt model performance due to excessive simplification.
* The right-hand rise in the curve shows that too much penalization (stronger regularization) reduces the model’s predictive ability.
* The vertical dotted lines show:  
  + The lambda that gives the lowest error (left line).
  + The largest lambda within 1 standard error of the minimum (right line), often chosen to simplify the model further without significantly increasing error.

#### **Ridge Model (Bottom Plot):**

* The curve shows a more gradual and smooth increase in deviance as log(λ) increases.
* There is a clear minimum on the left side, after which the model’s performance steadily declines with stronger penalization.
* Compared to Lasso, the curve is less flat near the minimum, suggesting that Ridge is more sensitive to the choice of λ.
* Ridge does not force coefficients to zero, but the plot still shows that regularization matters – too much of it leads to underfitting.

# **Conclusion**

This project used Lasso and Ridge logistic regression to predict depression in students based on sleep, diet, and academic workload. Both models performed well, with AUC scores above 0.82 and an accuracy of about 73%. Lasso helped highlight key risk factors by removing less important ones, while Ridge retained all variables with reduced influence. We found that unhealthy diets greatly increase the chance of depression, as they had high positive scores in our models (like 0.5 for Lasso). Good sleep lowers depression risk, with strong negative scores (like -0.3 for Lasso), acting like a shield. Academic pressure had a smaller effect, with low scores (like 0.1 for Lasso), meaning it matters less. These models offer useful tools for understanding and addressing mental health issues in college students.