Predicting Student Stress Levels Using Classification Techniques: Logistic Regression, LDA, and QDA

Group 1
Allana Coleman, Chonlachart Jeenprasom,
Tanchanok Sirikanchittavon, Sebin Nichlet

Florida International University
STA 6636 High Dimensional Data Analysis
April 25, 2025

1 Introduction

The prevalence of mental health challenges among university students has become a growing concern, particularly in high-stress academic environments. Stress, if left unaddressed, can lead to adverse outcomes, including academic underperformance, depression, and anxiety. In response to this issue, this study focuses on building predictive models to classify students' stress levels based on a combination of behavioral, psychological, and lifestyle variables. The primary goal is to apply and evaluate three classification algorithms which are Multinomial Logistic Regression, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA) in terms of their effectiveness at categorizing stress into three distinct levels: Low, Moderate, and High. These models were selected for their widespread use in high-dimensional classification tasks, with varying assumptions about data structure and distributions. The objective of the study is to compare the classification performance of three statistical learning methods. Therefore, early identification and intervention based on predicted stress levels can be a valuable tool in educational institutions.

2 Data Set

The dataset used in this study was retrieved from Kaggle, a publicly accessible data repository [3]. It comprises 760 responses from students on various attributes related to mental health and lifestyle. The dataset includes both categorical and continuous variables.

The categorical variables are Gender (Male, Female, and Others), Counseling Attendance

(Yes or No), Family Mental Health History (Yes or No), and Medical Condition (Yes or No). The continuous variables include Academic Performance (GPA), Study Hours per Week, Sleep Duration (hours per night), Physical Exercise (hours per week), Age (18–30), and Social Media Usage (hours per day).

Additionally, several variables are measured on a Likert scale from 1 to 5, including Family Support, Financial Stress, Peer Pressure, Relationship Stress, Diet Quality, Cognitive Distortions, and Substance Use, which were treated as continuous variables.

Data pre-processing involved several key steps: cleaning missing or inconsistent values, encoding categorical variables using dummy variables, and grouping the stress level variable into three categories (Low: 1–3, Moderate: 4–7, High: 8–10).

3 Exploratory Analysis

To understand the initial structure of the data, exploratory visualizations and statistical summaries were generated. Histograms and boxplots revealed skewed stress distributions by gender. Scatter plots were created between numeric predictors and stress levels, using simple linear smoothing for trends.

Distribution of Mental Stress Level

A histogram was created to visualize the distribution of students' mental stress levels (before categorization into Low, Moderate, and High). As shown in Figure 1, the distribution is moderately skewed with peaks around moderate stress levels.

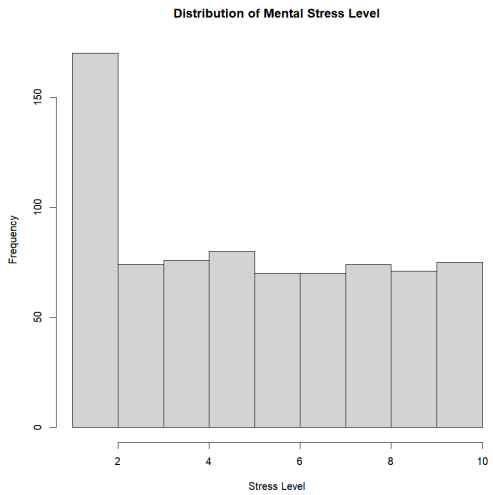

Figure 1: Distribution of Mental Stress Level (Original Scale)

ANOVA Analysis

An ANOVA (Analysis of Variance) was conducted to assess whether categorical variables have a statistically significant effect on students' stress levels. The results showed that none of the variables: Gender, Counseling Attendance, Family Mental Health History, and Medical Condition, were statistically significant at the 5% level.

- Gender: $F(2, 757) = 0.381, p = 0.683$
- Counseling Attendance: $F(1, 758) = 1.855, p = 0.174$
- Family Mental Health History: $F(1, 758) = 0.002, p = 0.989$
- Medical Condition: $F(1, 758) = 0.635, p = 0.426$

4 Methodology

This section details the statistical techniques used to classify students into low, moderate, or high stress categories. Each model was trained using an 80/20 train-test split and evaluated using a 10-fold cross-validation.

4.1 Multinomial Logistic Regression

The response variable in the dataset is the severity of stress classified into three categories (Low, Moderate, High). Multinomial Logistic Regression (MLR) predicts categorical outcomes by estimating the probability of class membership based on multiple independent variables. The model maximizes the conditional likelihood of observing the given class given the predictor values. MLR accommodates both binary (dummy) and continuous independent variables, making it appropriate for this dataset which includes mixed variable types.

The modeling strategies employed in this study follow standard approaches outlined in [1, 2, 4].

Mathematical Formulation for Multinomial Logistic Regression

Let K be the number of classes. The probability of observation i belonging to class k is modeled as:

$$P(Y_i = k \mid \mathbf{x}_i) = \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}_k}}{\sum_{l=1}^{K} e^{\mathbf{x}_i^\top \boldsymbol{\beta}_l}}, \quad \text{for } k = 1, \ldots, K$$

where $\boldsymbol{\beta}_k$ is the coefficient vector for class k, and $\mathbf{x}_i$ is the predictor vector for the i-th observation.

4.2 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) uses Bayes' theorem to calculate posterior probabilities for class membership based on predictor variables. It assumes that each class follows a multivariate Gaussian distribution and that all classes share a common covariance matrix. LDA seeks to maximize the separation between classes by modeling the joint likelihood of predictors and class labels.

Mathematical Formulation for Linear Discriminant Analysis

The classification rule in LDA is based on Bayes' theorem:

$$P(Y = k \mid \mathbf{x}) = \frac{\pi_k \, f_k(\mathbf{x})}{\sum_{l=1}^{K} \pi_l \, f_l(\mathbf{x})}$$

where:

- π_k is the prior probability of class k,
- $f_k(\mathbf{x})$ is the class-conditional density for class k,
- K is the total number of classes.

LDA further assumes that the class-conditional density $f_k(\mathbf{x})$ follows a multivariate Gaussian distribution:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

where:

- $\boldsymbol{\mu}_k$ is the mean vector of class k,
- Σ is the common covariance matrix shared across classes,
- p is the number of predictors.

4.3 Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) relaxes the assumption of LDA of equal covariance matrices across classes, allowing more flexibility at the cost of estimating additional parameters. Each class is modeled with its own covariance matrix, resulting in quadratic decision boundaries between classes.

Mathematical Formulation for Quadratic Discriminant Analysis

The discriminant function for class k in QDA is given by:

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \log(\pi_k)$$

where:

- $\boldsymbol{\mu}_k$ is the mean vector of class k,
- Σ_k is the class-specific covariance matrix,
- π_k is the prior probability of class k.

4.4 Model Evaluation and Comparison

The classification models were assessed using 10-fold cross-validation. Performance was measured primarily by overall classification accuracy:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

In addition to accuracy, confusion matrices were generated for each model to evaluate class-specific performance. Metrics such as sensitivity, specificity, and balanced accuracy were also computed to provide a more detailed understanding of model behavior across different stress levels.

5 Results

5.1 Model Performance

The classification accuracy of each model after applying 10-fold cross-validation is summarized in Table 1. Overall, the models demonstrated modest predictive ability, with LDA

achieving the highest average accuracy.

Table 1: Classification accuracy of each model

Model	Accuracy (%)
Multinomial Logistic Regression	35.1
Linear Discriminant Analysis	37.8
Quadratic Discriminant Analysis	32.5

5.2 Confusion Matrices

Confusion matrices for each model are shown in Tables 2, 3, and 4. They provide insight into how well each model distinguished between low, moderate, and high stress levels.

Table 2: Confusion Matrix for LDA

Predicted / Actual	Low	Moderate	High
Low	13	10	14
Moderate	9	11	11
High	26	24	33

Table 3: Confusion Matrix for Logistic Regression

Predicted / Actual	Low	Moderate	High
Low	13	10	14
Moderate	9	11	10
High	26	24	34

Table 4: Confusion Matrix for QDA

Predicted / Actual	Low	Moderate	High
Low	14	16	14
Moderate	14	12	13
High	20	17	31

5.3 Top Correlated Predictors

In addition to classification modeling, an exploratory correlation analysis was conducted. The top five positive and negative predictors correlated with mental stress levels were identified. A bar chart summarizing these correlations is shown in Figure 2.

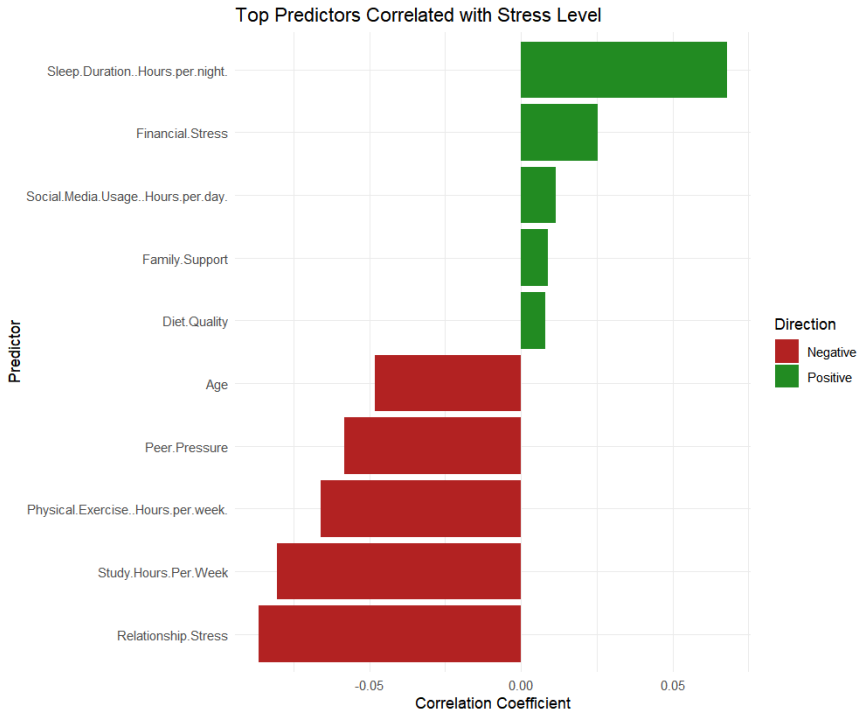

Figure 2: Top Predictors Correlated with Mental Stress Level

6 Discussion

Initial models showed inflated performance due to the inclusion of the Mental.Stress.Level variable in both predictors and the target (data leakage). After correcting for this issue, accuracy dropped significantly to more realistic values around 26–38%, highlighting the critical importance of preventing data leakage in machine learning pipelines.

Multinomial Logistic Regression and LDA performed similarly, while QDA showed lower accuracy due to multicollinearity and sensitivity to the small sample sizes in each stress category. Although the dataset originally included stress coping mechanisms as a multi-valued categorical variable, this was excluded during modeling to reduce dimensionality and avoid instability across cross-validation folds. This exclusion likely contributed to information loss, as coping strategies could have been informative predictors if appropriately encoded.

The dataset used in this study was synthetically generated for illustrative purposes rather than collected from real-world student populations. Consequently, the relationships among predictors and the outcome variable were relatively weak. Psychological stress is influenced by complex, nonlinear interactions and latent variables not easily captured through simple survey questions, which explains the modest predictive performance observed across all models. Some observed correlations, such as the negative relationship between relationship stress and mental stress, likely resulted from artifacts inherent to the simulated data rather than true behavioral patterns.

Exploratory analysis revealed that factors such as lower sleep duration, higher financial stress, and greater peer pressure were among the strongest predictors of mental stress, although their effects were relatively small. These results suggest detectable patterns that future studies could build upon.

Future improvements could include feature selection methods (e.g., LASSO), dimensionality reduction techniques (e.g., PCA), or ensemble learning models to better capture nonlinearities and interactions. Collecting richer, real-world longitudinal data, and applying hierarchical or mixed-effects models, could further enhance the predictive ability for mental stress classification in academic populations.

7 Conclusion

This study explored the classification of student mental stress levels using Multinomial Logistic Regression, Linear Discriminant Analysis, and Quadratic Discriminant Analysis. After correcting for data leakage and simplifying the feature space, classification accuracies were modest, with the best model achieving approximately 38% accuracy. Weak correlations between predictors and stress levels, along with inconsistencies in variable relationships, highlight the challenges of working with synthetically generated datasets. Future work could incorporate more advanced feature selection techniques, nonlinear modeling approaches, and richer real-world data sources to improve classification performance.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, second edition, 2009.

[2] Gareth James, Trevor Hastie, Daniela Witten, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, second edition, 2021.

[3] Salahuddin Ahmed Shuvo. Student mental stress and coping mechanisms, 2022. Accessed: April 01, 2025.

[4] Wensong Wu. Lecture notes for sta6636 high dimensional data analysis, 2025. Unpublished course notes, Florida International University.