Predicting Student Stress Risk Using R

Name: Daniel Habtu Gebrai

Date: 13/7/2025

Introduction

Student mental health has become a global concern, with academic pressure, financial stress, and social isolation contributing to increasing rates of depression and burnout. This project analyzes a dataset of students to identify and predict depression risk using logistic regression, supported by interactive visualizations through a Shiny web app.

The dataset includes self-reported data from students aged 15–40, covering variables like academic pressure, study satisfaction, sleep duration, suicidal ideation, and financial stress. The objective is to build a predictive model and create a dynamic app to help users self-assess their stress risk and receive evidence-based recommendations.

By visualizing student risk profiles, we can:

- Detect early warning signs of mental health deterioration.
- Support intervention planning for educators and health professionals.
- Empower students with awareness and tools for self-care.

This solution bridges data science and public health through actionable, interactive tools.

Key User Personas and Targets

The project is designed for three primary user types:

Students

Aspired Action: Self-assess stress risk and reflect on study-life balance.

Target: Increase awareness of mental health and seek timely help.

University Counsellors

Aspired Action: Use risk patterns to identify at-risk students.

Target: Provide targeted counseling and mental health support.

Researchers

Aspired Action: Analyze variables that predict student stress level.

Target: Study behavioral and academic predictors to improve policies and curricula.

Data Cleaning

The data cleaning script focused on preparing the raw data for analysis and modeling.

Steps Performed:

- **Library Imports:** dplyr, readr
- **Column Renaming:** Simplified column names for ease of use.
- **Column Selection:** Kept only relevant predictors (e.g., Gender, Age, Academic_Pressure, Suicidal_Thoughts).
- **Filtering:**
  - Removed rows with missing values in critical variables.
  - Filtered for students aged 15–40 to focus on higher education populations.

Output: A cleaned dataset (student_depression_cleaned.csv) containing relevant, complete observations.

Data Scoring and Visualization

Scoring was introduced during the visualization stage using the mutate() function in dplyr. The scores were assigned to convert categorical variables into numeric risk values.

New Variables Created:

- Gender_Score: Female = 0, Male = 1
- Suicidal_Score: Yes = 3, No = 0
- Family_History_Score: Yes = 2, No = 0
- Sleep_Score: Mapped from sleep duration
- Dissatisfaction_Score: 5 - Study Satisfaction
- Academic_Score: Academic Pressure * 1.5
- Financial_Score: Financial Stress * 2

Visualizations Developed:

- **Boxplots:** Compared risk scores by depression status. Higher scores in academic pressure, dissatisfaction, and suicidal thoughts were observed in depressed students.
- **Correlation Heatmap:** Showed relationships among numeric variables. Financial and academic pressure correlated with depression.
- **Pairwise Plots (GGally):** Helped identify variable clusters related to depression.
- **Parallel Coordinates Plot (Plotly):** Illustrated how high-risk individuals scored across multiple dimensions.

Shiny App Implementation

The Shiny app allows users to explore their own stress risk interactively.

UI Components:

- Age slider
- Dropdowns for categorical inputs: Gender, Academic Pressure, Study Satisfaction, Sleep Duration, Suicidal Thoughts, Study Hours, Financial Stress, Family History

Server Logic:

- A logistic regression model predicts depression probability based on scored variables.
- Risk percentage is calculated and color-coded.
- A personalized recommendation is provided based on predicted risk.

Model Specification:

```
model <- glm(Depression ~ Gender_Score + Age + Academic_Score + Dissatisfaction_Score +
Sleep_Score + Suicidal_Score + Study_Hours +
Financial_Score + Family_History_Score,
data = data, family = binomial)
```

Outputs:

- Risk percentage
- Risk category (Low, Moderate, High, Critical)
- Color-coded bar plot
- Tailored mental health recommendation