Machine Learning Approaches to Predicting Depression in University Students: A Comparative Analysis of Logistic Regression, LASSO, and Random Forest

Xinyu Wang

Department of Mathematics, New York University, New York, 10003, United States

xw2875@nyu.edu

Abstract. Depression is a growing mental health concern among university students and can harm academic performance and well-being. This study analyzes data from 27,901 university students to explore key risk factors and build models to predict depression. We focus on variables such as academic pressure, financial stress, sleep duration, study satisfaction, family history of mental illness, age, gender, and grade point average. Three machine-learning methods—logistic regression, LASSO, and random forest—were applied and compared. The results show that academic pressure, financial stress, and short sleep are the strongest predictors of depression, with family history and low study satisfaction also playing important roles, while age, gender, and grades had smaller effects. All three models performed strongly, with area-under-the-curve (AUC) values around 0.86 and showed good calibration. Logistic regression and LASSO achieved nearly the same accuracy as random forest, making them easier to use and explain in real settings. These findings highlight practical steps for schools and health services: programs that lower academic pressure, provide financial help, and improve sleep habits can reduce depression risk. Early identification and timely support may help prevent serious problems. This research shows that simple machine-learning models can guide effective and affordable mental-health strategies for students.

Keywords: Depression, Machine Learning, Risk Factors, Predictive Models, Data Visualization.

1. Introduction

Depression is a common mental disorder that is becoming increasingly prevalent and catching people's attention in the public health sector. It is not only a medical or psychological concern but also a social issue that significantly affects individuals and communities. Depression is often characterized by several key symptoms, such as persistent sadness or low mood, a noticeable loss of interest or enjoyment in usual activities, heightened agitation or slowed behavior, constant fatigue or lack of energy, diminished self-confidence and self-esteem, difficulty concentrating or making decisions, and in severe cases, recurring thoughts of death, suicidal ideation, or suicide attempts [1]. These diagnostic criteria align with the definitions provided by DSM-5 and WHO guidelines. Because of the symptoms, students become an especially fragile group to the occurrences and consequences of depression. Meanwhile, schools are becoming highly concerned about the issue of depression among their students and are putting a high amount of money and energy in the preventions and treatments on depression.

Research indicates that depression has a clear and detrimental impact on students' academic performance, and there are significant differences in achievement outcomes among students experiencing varying levels of depressive symptoms. [2]. Cahuas, He, Zhang, and Chen discovered that patterns of vigorous physical activity, along with sleep-related variables, serve as significant predictors of depression levels, suggesting that both exercise habits and sleep quality play an important role in influencing an individual's risk of experiencing depressive symptoms [3]. However, physical activity and sleep duration might not be the only factors that greatly affects students' depression levels. Cassady, Pierson, and Starling demonstrated that academic anxiety is a significant predictor of depression among college students. They further proposed that heightened levels of perceived stress or threat in response to academic challenges could function as an early warning indicator, helping educators and researchers identify learners who may be at increased risk of

academic failure and the development of depressive symptoms [4]. Machine learning is a useful tool that can be integrated into the analysis of depression to try to predict its occurrence and severity. Sawangarreerak and Thanathamathee' random forest with a more effective method in handling imbalanced data turns out to effectively predict university student depression [5]. Still, other machine learning techniques may also work effectively and predict accurately. Therefore, it would be helpful to utilize different machine learning methods and compare their results to figure out which ones are better at predicting students' depression levels.

Since depression is a complicated illness, a comprehensive analysis of this problem necessitates consideration from multiple dimensions. The nonpathological related factors of students' depression mainly fell into four categories: biological factors, personality and phycological state, college experience, and lifestyle [6]. Therefore, considering factors and variables that come from different aspects is essential. This paper aims to explore the relationship between depression and various factors such as age, gender, academic pressure, sleep duration, financial stress, and family history of mental illness. The paper tries to identify the key factors that are significantly related to depression. Then the paper hopes to use three machine learning methods – logistic regression, LASSO regression, and random forest – to try to predict depression. The results of the three methods would be evaluated and compared. Therefore, by integrating and comparing multiple machine learning approaches, this study not only provides a more rigorous understanding of the factors influencing student depression but also helps build more accurate prediction models. Such models can greatly reduce the reliance on costly manual screening, saving both time and resources for schools and public health institutions. More importantly, through earlier detection and more targeted prevention strategies, they hold the promise of improving treatment outcomes and ultimately contributing to the broader goal of enhancing human well-being and social development.

2. Methods

2.1. Data Overview

After cleaning and preprocessing, the dataset used in this study contains information from 27,901 students, including whether each individual screened positive for depression. Each entry includes more than 20 variables ranging from demographic background, perceived stress levels, to other mental health indicators. The original data came from Kaggle, an online community that provides datasets for research and data science applications.

The data-preparation process can be summarized in the following steps: converting Yes/No responses into binary values, transforming categorical variables into numerical factors, and organizing sleep duration into ordered categories. These adjustments will make the future statistical evaluation and machine learning applications easier and more efficient.

2.2. Attributes

The key variables studied in this paper can be grouped into four different categories, each capturing a distinct aspect of potential influence. The first category includes demographic characteristics such as age and gender. These characteristics help illustrate how personal background might be related to students' depression. The second category includes academic factors, for example cumulative grade point average, satisfaction with academic performance, and academic pressure. These factors tell us how educational experiences may affect students' mental health. The third category includes lifestyle and external factors, such as working hours, financial stress, and sleeping hours. This category helps with the analysis of how everyday activities and financial conditions influence depression risks among students. The fourth category involves clinical and health-related variables, such as family history of depression and past experiences of suicidal thoughts. Having all these variables together, this paper provides a multidimensional analysis of the various factors associated with students' depression.

2.3. Exploratory Analysis

The patterns within the data and the relationships between variables were studies during the exploratory analysis phase. Visualizations and descriptive summaries were used in this step. Histograms were graphed to examine the distributions of continuous variables, and proportional bar charts were used to compare categorical variables. Visualizations can reveal variations in depression rates across the key factors. Exploratory analysis can highlight trends and imbalances, which helps identify potential predictors, validate assumptions, and guide subsequent modeling decisions. It also helps sdetect unusual patterns that may be considered and handled.

2.4. Algorithms

The three machine learning approaches that are applied in this paper are logistic regression, random forest, and LASSO regression.

Logistic regression is a statistical method used to predict binary outcomes. In this study, the outcomes are either having or not having depression. It models the log-odds of depression as a linear function of the input variables and estimates the probability of the outcome. After exponentiating the coefficients, odds ratios can be derived. These odds ratios indicate how each factors increases or decreases the likelihood of depression and quantify their influences [7].

Random forest is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions to improve accuracy and robustness. Each tree is built using a random sample of the data and a random subset of features, which helps minimize overfitting and enhances generalization. For classification tasks, the final prediction is determined by majority voting across all trees. The model also estimates variable importance, identifying which predictors have the strongest impact on the outcome [8].

The Least Absolute Shrinkage and Selection Operator (LASSO) is a regularized regression technique that performs both variable selection and coefficient regularization. It applies a penalty based on the absolute values of the coefficients, which forces some of them to shrink to zero—effectively eliminating less relevant predictors. In the context of logistic regression, LASSO helps avoid overfitting, handles multicollinearity, and often improves predictive accuracy. This makes it especially valuable when working with datasets that contain many correlated variables, as it produces a more interpretable and sparser model without compromising performance [9].

2.5. Evaluation Metrices

Complementary metrics are used to evaluate the performance of each model. The Area Under the ROC Curve (AUC) tells us how well each model distinguishes between students with and without depression. A higher AUC value reflects a better overall prediction. Metrics like accuracy, sensitivity, specificity, positive predictive value, and negative predictive value are derived from the confusion matrix. These metrics offer deeper insight into classification performance. These statistics help assess how accurate and effective each model is in identifying true cases of depression. Calibration analysis compares predicted probabilities against observed outcomes, ensuring that risk estimates are not only discriminative but also reliable in magnitude. Additionally, the Brier score measures overall probability accuracy, with lower scores reflecting better calibration and sharper predictions. By combining these metrics, the evaluation gives a comprehensive picture of model effectiveness, highlighting both strengths and limitations across logistic regression, random forest, and LASSO [10].

3. Results

3.1. Logistic Regression

Figure 1 shows the top adjusted predictors by effect size. Academic pressure is the strongest, followed by financial stress, sleep duration (linear), family history of mental illness, and study satisfaction; work study hours, age, gender, and cumulative grade point average have smaller effects.

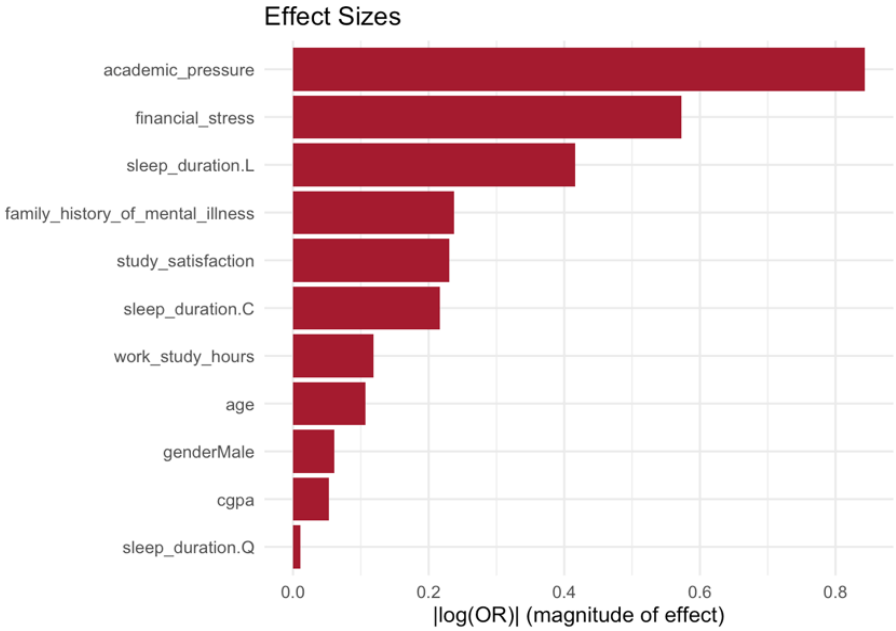

Figure 1. Logistic Regression effect sizes (Picture credit: Original)

3.2. Random Forest

In Figure 2, the model highlights academic pressure as the most influential predictor, with financial stress and age emerging as the next most influential factors, with moderate impacts. Mean Decrease Accuracy shows how much worse the prediction gets if the variable is removed, and Mean Decrease Gini shows how useful the variable is for splitting data during tree construction.

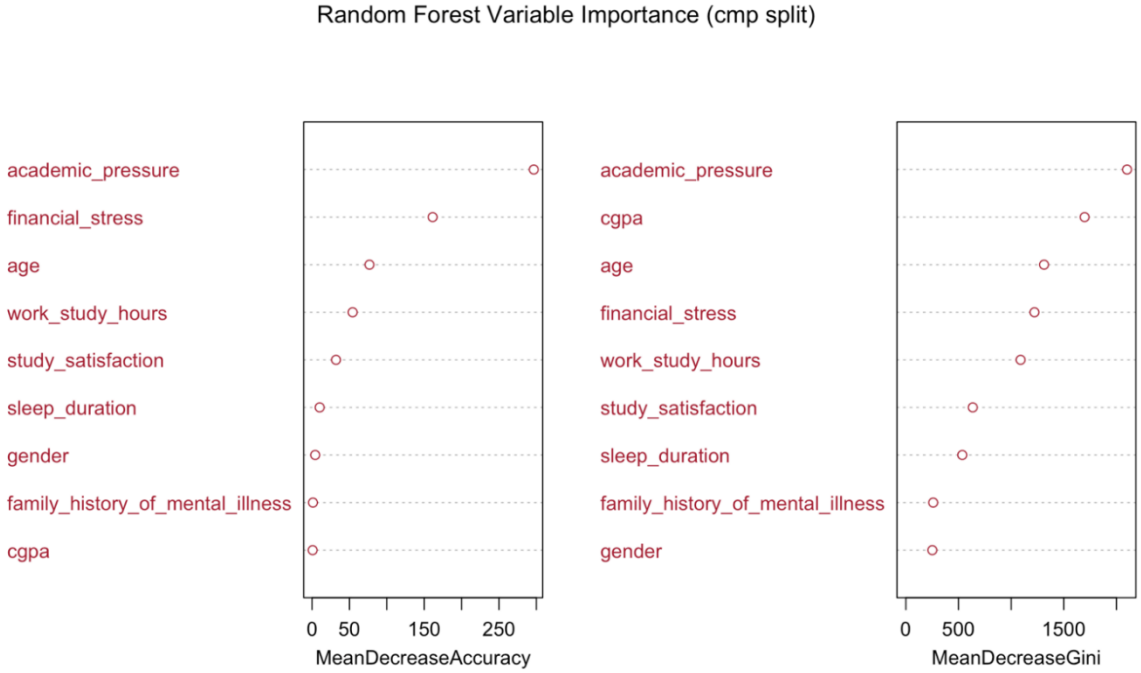

Figure 2. Random Forest variable importance (Picture credit: Original)

3.3. LASSO

In Figure 3, the LASSO model identifies academic pressure, financial stress, and short sleep as the strongest positive predictors of depression, with moderate contributions from family history of mental illness and study satisfaction. Smaller effects appear work-study hours, age, gender, and cgpa, indicating these factors add little to the overall risk compared with the leading stress- and sleep-related variables.

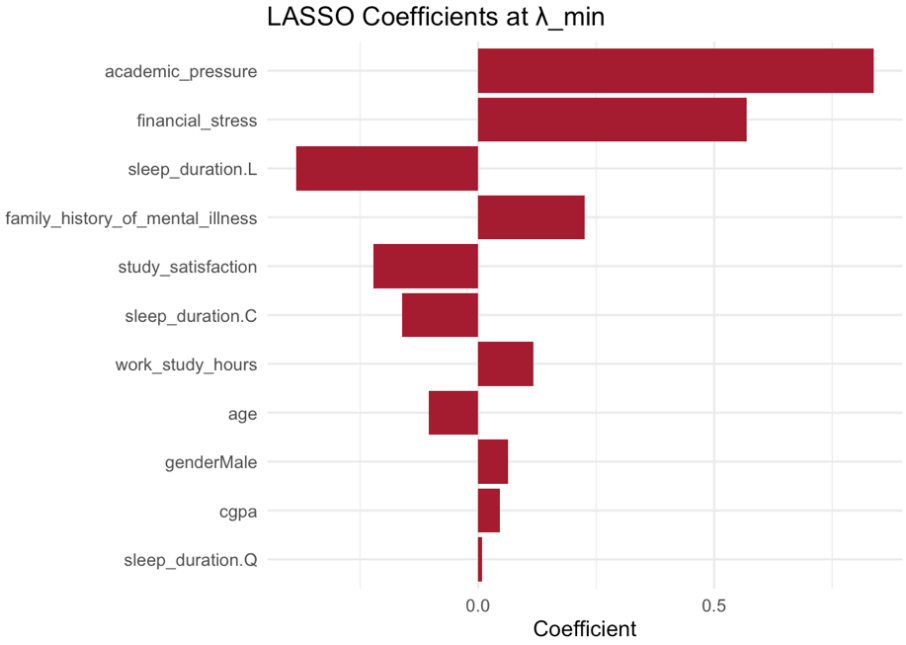

Figure 3. Random Forest variable importance (Picture credit: Original)

3.4. Comparisons

In Figure 4, The ROC comparison shows that all three models—Logistic Regression, Random Forest, and LASSO—perform similarly well in distinguishing depressed from non-depressed cases. Their AUCs are nearly identical (≈0.86), indicating strong overall discrimination. The curves overlap across the false-positive range, with Random Forest only marginally lower (AUC = 0.85). This suggests that simpler linear models (Logistic, LASSO) achieve almost the same predictive power as the more complex Random Forest**,** making them attractive for interpretable modeling without sacrificing accuracy.

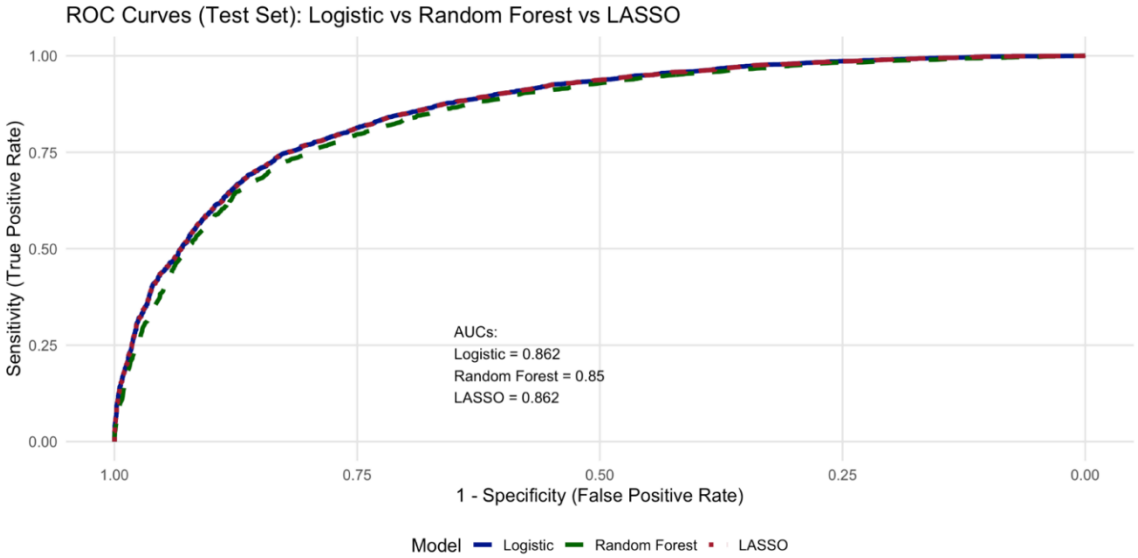

Figure 4. ROC curves (Picture credit: Original)

In Figure 5, The calibration plot shows that all three models—LASSO, Logistic Regression, and Random Forest—are well-calibrated on the test set**.** Across the full range of predicted probabilities, their red calibration curves track closely with the diagonal 45-degree reference line, indicating that predicted risks match observed depression rates in each decile. Minor deviations appear only at the highest probabilities, where data are sparser, but overall, the agreement is strong. This suggests that each model not only discriminates well but also provides reliable probability estimates, making their risk predictions trustworthy for practical interpretation.

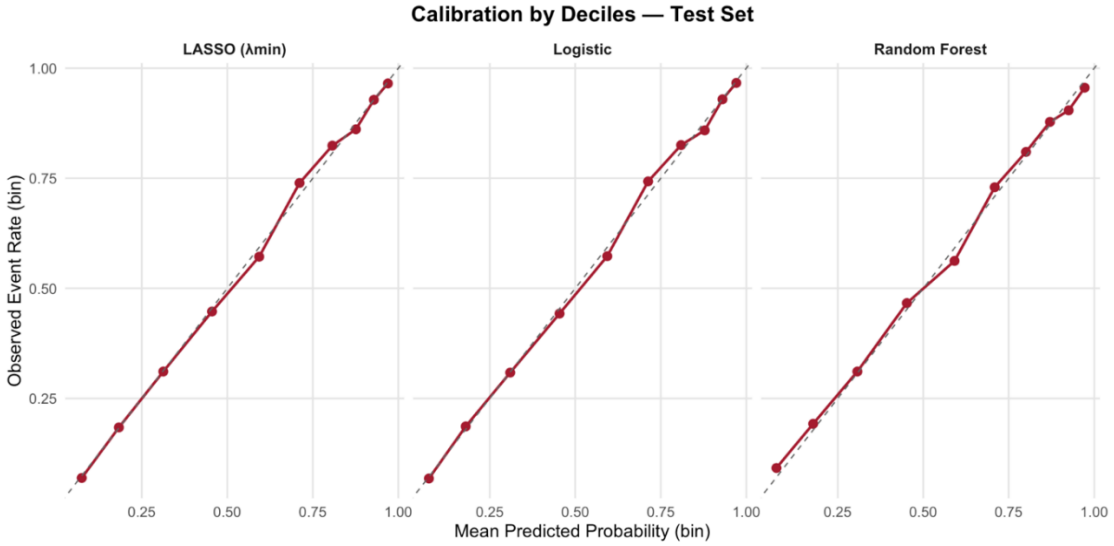

Figure 5. Calibration by deciles (Picture credit: Original)

4. Summary

This study shows that data analysis and machine-learning methods can help us better understand and predict depression in university students. Using a large dataset of 27,901 students, we tested three models—logistic regression, LASSO, and random forest—to find the strongest signs of depression. Across all three models, academic pressure, financial stress, and short sleep stood out as the main risk factors. Family history of mental illness and low study satisfaction also mattered, while age, gender, and grade point average had smaller effects.

When we compared the models, we found that more complicated methods do not always perform better. Random forest is powerful and flexible, but logistic regression and LASSO reached almost the same level of accuracy, with area-under-the-curve (AUC) values around 0.86. All three models were well calibrated, meaning their predicted risks matched real outcomes. This is important because simpler models are easier to explain and to use in real school or health programs.

These results give schools and health workers practical ideas for action. Programs that lower academic pressure, provide financial help or advice, and encourage good sleep habits could reduce depression risk. Schools can use these models to spot students who may need support earlier and to give help before problems grow more serious. Doing so may lower the long-term personal and social costs of depression.

Despite its contributions, certain limitations should be acknowledged. This research does have limits. The data come from one point in time, so it cannot show cause and effect. Other factors, such as friendships, personality, or social support, were not included. Future studies could add long-term tracking or new types of data, such as information from apps or daily digital records, to build even stronger prediction tools.

Overall, this project shows that combining simple but strong machine-learning methods with well-chosen risk factors can make depression screening more accurate and affordable. By helping schools detect problems sooner and plan targeted support, these methods can improve students' mental health and contribute to the wider effort to reduce depression in society.

References

[1] Paykel E S. Basic concepts of depression. Dialogues in Clinical Neuroscience, 2008, 10 (3): 279-289.

[2] Khurshid S, Parveen Q, Yousuf M I, Chaudhry A G. Effects of depression on students' academic performance. Science International, 2015, 27 (2): 1619-1624.

[3] Cahuas A, He Z, Zhang Z, Chen W. Relationship of physical activity and sleep with depression in college students. Journal of American College Health, 2020, 68 (5): 557-564.

[4] Cassady J C, Pierson E E, Starling J M. Predicting student depression with measures of general and academic anxieties. Frontiers in Education, 2019, 4: 11.

[5] Sawangarreerak S, Thanathamathee P. Random Forest with sampling techniques for handling imbalanced prediction of university student depression. Information, 2020, 11 (11): 519.

[6] Liu X Q, Guo Y X, Zhang W J, Gao W J. Influencing factors, prediction and prevention of depression in college students: a literature review. World Journal of Psychiatry, 2022, 12 (7): 860.

[7] Sperandei S. Understanding logistic regression analysis. Biochemia Medica, 2014, 24 (1): 12-18.

[8] Breiman L. Random forests. Machine Learning, 2001, 45 (1): 5-32.

[9] Ranstam J, Cook J A. LASSO regression. Journal of British Surgery, 2018, 105 (10): 1348-1348.

[10] Dinga R, Penninx B W J H, Veltman D J, Schmaal L, Marquand A F. Beyond accuracy: measures for assessing machine learning models, pitfalls and guidelines. bioRxiv, 2019: 743138.