# Predicting Student Depression using Machine Learning: A Comparative Study of Logistic Regression and Random Forest

Yijin Jiang*

Department of mathematics, London School of Economics and Political Science, London, WC2A 2AE, United Kingdom

*Corresponding author: y.jiang82@lse.ac.uk

**Abstract.** Depression is a common mental health concern among university students, typically caused by academic and social demands. The purpose of this study is to investigate the efficacy of machine learning techniques in detecting depression in its early stages. To optimize model parameters, this research used a Kaggle dataset with 502 participants, rigorous data preprocessing, and a 10-fold cross-validation strategy. Two predictive models were created: logistic regression with elastic net regularization and random forest model. The results reveal that the logistic regression model obtained an accuracy of 98%, beating the random forest model's 92% accuracy. At the same time, feature importance analysis highlighted academic pressure and suicidal ideation as significant predictors. These results highlight data-driven approaches as likely to improve early diagnosis and targeted intervention of mental health issues in universities. Thus, the research informs depression understanding in student communities, providing a comparative perspective to the effectiveness of various predictive models for guiding preventative and support measures.

**Keywords:** Depression; logistic regression; random forest; prediction model.

## 1. Introduction

Depression is a common mental disorder that affects academic performance, social interactions, and overall quality of life. It is best identified by a persistent low mood, a loss of interest in enjoyable activities, persistent tiredness, and reduced cognitive ability. Among college and university students, who are constantly under immense academic and interpersonal pressures, depression is becoming an emergent public health problem with negative consequences for their physical and psychological well-being, academic performance, and overall well-being [1]. Recent research found a considerably high incidence of depression in this group. For example, a study by Ahmed et al. found that almost one-third of students suffer from depression, with a weighted overall prevalence of 30.6%, as against only 9% found in the general population [2]. Several stress factors, including heavy course loads and economic pressures, significantly add to the development of depression risks [3]. Further, the disorder can lead to serious complications, including suicidal ideation [4]. It is therefore critical to use data-driven analytical methods to fully examine contributing factors towards developing effective intervention guidelines to combat student depression and its harmful consequences.

Academic stress ranks as one of the most frequently cited factors for depression because students must cope with heavy course loads, extended study sessions, and high performance expectations. Academic accomplishment comes with chronic stress, anxiety, and test anxiety, all of which have been strongly ascertained to increase the risk of depression [5]. Moreover, these negative stressors not only manifest as psychological burdens but also cause significant physiological changes. For instance, stress is identified to increase serum leptin and cortisol levels [6]. These physical changes have the potential to exacerbate academic performance, thereby bringing about a vicious cycle of stress and, if untreated, to increase depression development.

In addition, financial stress is a significant factor affecting mental well-being for students [7]. The cost of student loans, living, and tuition is a huge economic burden to many students, which intensifies psychological stress. Part-time jobs are needed for some students to cover living conditions alongside education, thus exacerbating stress and negatively influencing well-being. Moreover, volatile economic circumstances can erode access to fundamentals such as housing and food, thus evoking a

deep sense of insecurity and negatively influencing well-being. A study by Silvana et al. confirmed depressive symptoms to be evident at higher levels for students from low-income families [8]. Financial stress can bring social withdrawal behaviors, as limited-resource students find themselves unable to participate in social and academic events, which entail economic expenditure, thus exacerbating feelings of isolation and propagating an increased risk for depression. This finding is supported by research utilizing multilevel modeling and functional regression [9].

Sleep deprivation is likely to be another cause of student depression. The conflict between different academic requirements and enjoyments tends to cause poor and unpredictable sleeping schedules for students. Research has postulated that sleep deprivation greatly increases the likelihood for depression, as with survey logistic regression modeling [10]. Unhealthy eating habits are likewise associated with poor psychological health. For instance, those who regularly consume fast food, sweets, and other high-calorie, high-fat foods show symptoms of depression [11]. Additionally, a history of mental illness within families is likely to predispose students to be at larger risks for depression, as with Milne et al.'s study on a linear regression model, which found there to be a well-established association between the two [12].

These research efforts will attempt to identify how academic, personal, and lifestyle factors are related to depression for college students. Using a dataset of 502 students, it will describe the most critical factors of depression for student populations and compare them with regards to how critical these factors are. In order to compare how well various sets of factors predict depression, statistical modeling techniques, such as logistic regression, will be employed as well as machine learning techniques, such as random forests. The ultimate purpose of these research efforts is to make data-driven distinctions that can be employed to inform both construction of mental health programs as well as ancillary systems of care for college education. By identifying key factors associated with depression risk, the study aims to inform the creation of more effective prevention and intervention strategies, thereby helping to improve students' mental health outcomes.

## 2. Methods

### 2.1. Data Source

Data for this study came from Kaggle and included replies from 502 participants. The dataset contains demographic, educational, and lifestyle characteristics related to mental health. It is in CSV format and does not contain any missing values. The dataset facilitates quantitative investigation of factors linked with student depression by providing explicitly specified variables for statistical evaluation and model creation, so establishing a replicable foundation for empirical research and future objective analysis.

### 2.2. Variable Introduction

The original dataset comprised 11 variables, with their names and descriptions presented below in table 1. Categorical variables were standardized in order to make statistical analysis easier. "Yes" was encoded as 1 and "No" as 0, converting the binary variables (Depression, Suicidal Thoughts, and Family History of Mental Illness) to numeric values. Multi-category variables (such as gender, sleep duration, and dietary habits) were transformed into factors.

**Table 1**. Variables description

| Variables | Type | Explanation |
|---|---|---|
| Gender | Categorical | Categorized as "Female" and "Male" |
| Age | Numeric | Age of participant in years |
| Academic Pressure | Numeric | Scale 1 (low) - 5 (high) |
| Study Satisfaction | Numeric | Scale 1 (low) - 5 (high) |
| Sleep Duration | Categorical | Categories: Less than 5 hours, 5-6 hours, 7-8 hours, More than 8 hours |
| Dietary Habits | Categorical | Categories: Healthy, Moderate, Unhealthy |
| Have you ever had suicidal thoughts? | Categorical | 0 = No, 1 = Yes |
| Study Hours | Numeric | Hours per day |
| Financial Stress | Numeric | Scale 1 (low) - 5 (high) |
| Family History of Mental Illness | Categorical | 0 = No, 1 = Yes |
| Depression | Categorical | 0 = No, 1 = Yes |

### 2.3. Method Introduction

This study uses logistic regression and random forest models to investigate the risk factors associated with depression and to enhance the accuracy of depression prediction. The dataset is split into a training set (80%) and a testing set (20%) to preserve objectivity. The training set is used for model fitting, while the testing set is reserved for assessing generalization performance. Cross-validation applied to the training set is employed to make models more stable and to facilitate model optimization and hyperparameter tuning.

Logistic regression is applied to identify and quantify the effects of potential risk factors on the likelihood of depression. The model estimates the probability $P$ of depression based on multiple independent variables, and is expressed as:

$$P = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)} \quad (1)$$

Where $\beta_0$ is the intercept term, and $\beta_1, \beta_2, \ldots, \beta_m$ are regression coefficients for predictor variables $x_1, x_2, \ldots, x_m$. By applying the logit transformation, the model can be rewritten as:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \quad (2)$$

This model is capable of direct interpretation of how all factors influence depression odds, and hence it is an appropriate model for detecting statistically significant psychological and lifestyle factors contributing to mental health status.

A random forest model is trained to detect interactions and nonlinear relationships between variables. In order to make it more resilient, as an ensemble learning technique, this model constructs many decision trees and sums up their output. Feature importance scores are computed to identify which predictors are most impactful.

A confusion matrix is used to evaluate the model. To evaluate each model's capacity for discrimination, the area under the curve (AUC) and receiver operating characteristic (ROC) curves are also calculated. Both interpretability and prediction accuracy are taken into account when comparing the performance of random forest with logistic regression.

By analyzing the results, this study determines the model that best predicts depression risk while offering meaningful insights into the contributing factors.

## 3. Results and Discussion

### 3.1. Descriptive Analysis

Figure 1 shows the distributions of key variables, including Age, Gender, Study Hours, and several lifestyle-related measures. Age is spread roughly from late teens to early thirties, with each category

relatively balanced. Academic Pressure, Financial Stress, and Study Satisfaction vary widely, suggesting diverse student experiences.

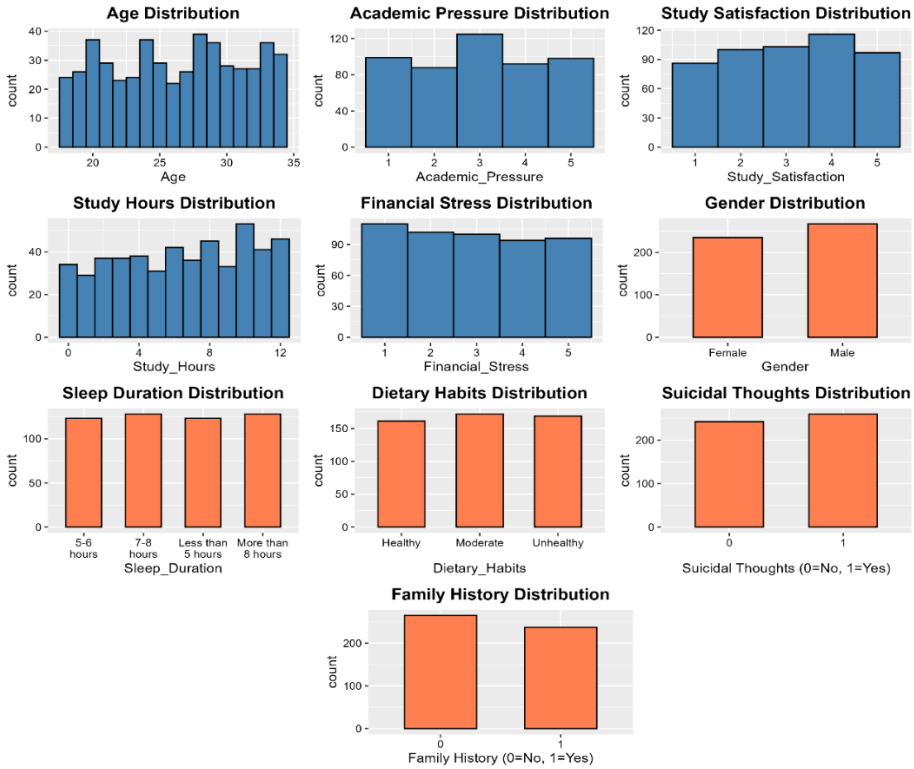


**Fig. 1** Distribution of all dependent variables

Figure 2 is the Spearman correlation matrix highlighting Depression. Depression correlates positively with Academic Pressure (0.48), Suicidal Thoughts (0.47), and Financial Stress (0.30), indicating higher depression risk under greater stress. Conversely, it has a moderate negative correlation with Study Satisfaction (−0.29), suggesting that increased satisfaction may reduce depression levels. Other variables, such as Sleep Duration and Dietary Habits, show weaker correlations but remain potentially relevant for predictive modeling. The Gender variable was removed during data preprocessing because it did not show a significant impact on depression in preliminary analyses, thereby ensuring greater model stability and predictive accuracy.

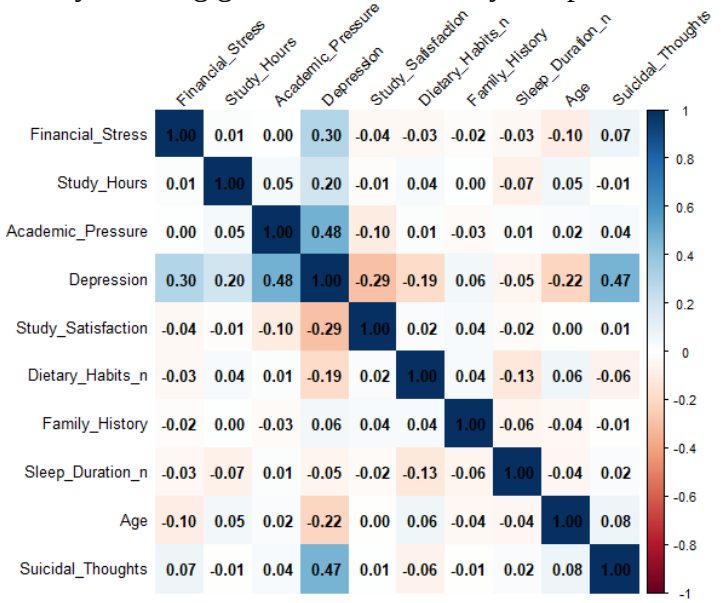


**Fig. 2** Variables Correlation

### 3.2. Logistic Regression Model

This study implemented an Elastic Net logistic regression model to investigate factors associated with depression. The dataset was split into a training set (80%) and a testing set (20%). A 10-fold cross-validation was used on the training data to optimize the Elastic Net parameters $\alpha$ (mixing percentage) and $\lambda$ (penalty strength).

The Elastic Net approach combines L1 (Lasso) and L2 (Ridge) penalties, allowing it to both shrink coefficients and set some of them to zero if needed. Cross-validation selected $\alpha$ and $\lambda$ that maximized the area under the ROC curve (AUC) on the training folds. Figure 3 illustrates how the model's cross-validated AUC varied across different values of and $\lambda$.

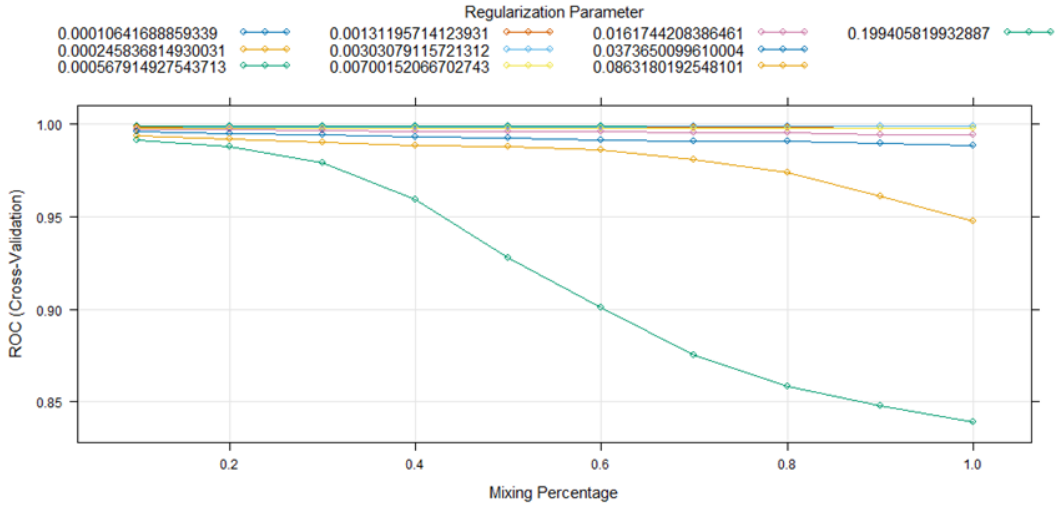


**Fig. 3** Regularization Parameter Plot (Alpha-Lambda Plot)

Figure 4 depicts these coefficients graphically. Positive coefficients indicate risk factors that increase the log-odds of depression, while negative coefficients indicate protective actors. Variables are sorted by the absolute size of their contribution to the log-odds of depression. The coefficients corresponding to Suicidal Thoughts to Age are 3.7581, 3.6014, 2.0846, 1.6738, 0.7292, −0.8407, −1.3181, −2.1690, and −2.2955, while intercept is -0.1568.

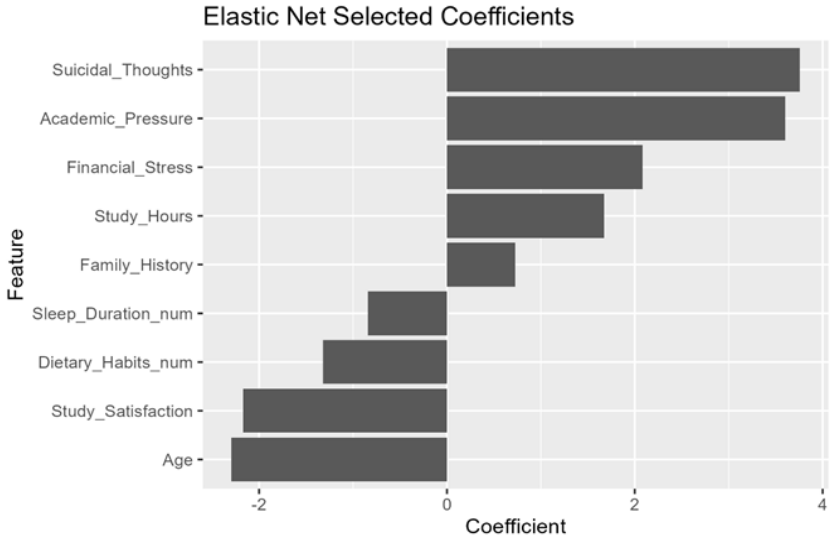


**Fig. 4** Elastic Net Selected Coefficients Bar Chart

The trained model was applied to the hold-out test set to assess its performance on unseen data. Figure 5 presents the confusion matrix and its accuracy is 0.9800. It can be seen that 48 participants actually did not have depression and were not predicted to have depression; 2 participants actually

did not have depression but were predicted to have depression; 50 participants actually have depression and were predicted to have depression; and no patients actually have depression but were predicted to not have depression.

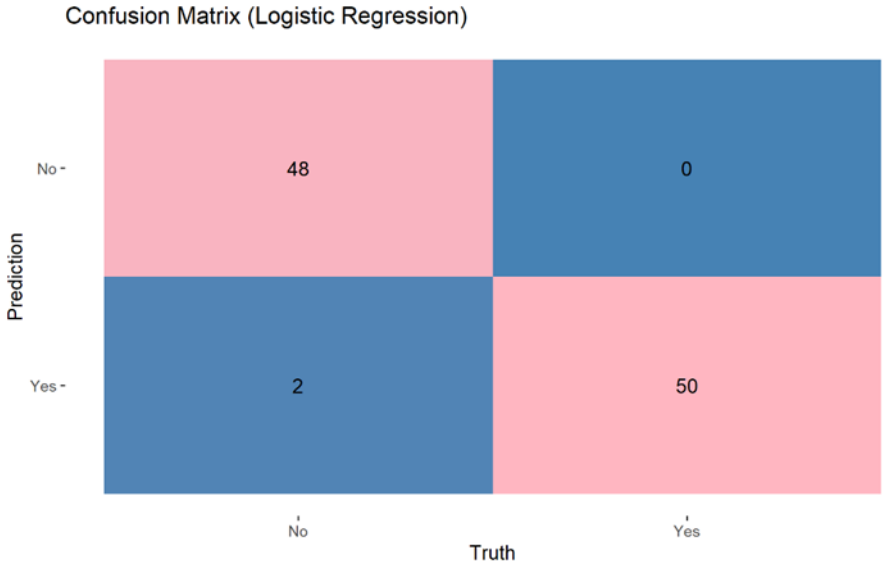


**Fig. 5** Confusion Matrix of logistic regression model

### 3.3. Random Forest Model

The training of random forest model is under the same 10-fold cross-validation scheme applied to the Elastic Net model. The same training set (80% of the data) was used for model fitting, and the remaining 20% was reserved for testing. To ensure the reproductivity of the results, this article sets the random seed to "456". ROC was used to select the optimal model using the largest value. The final value used for the model was mtry = 3.

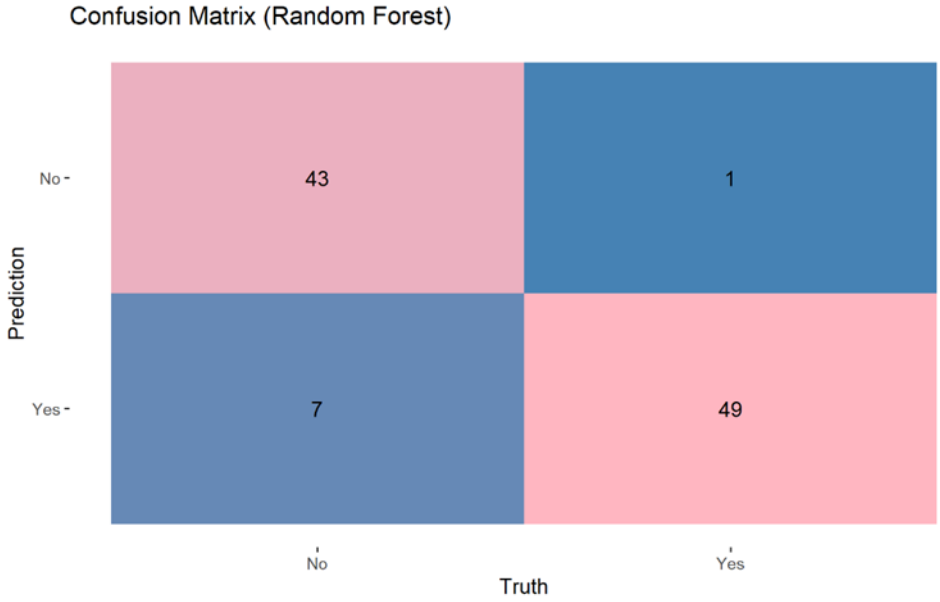


**Fig. 6** Confusion Matrix of random forest model

Figure 6 presents the confusion matrix and its accuracy is 0.9200. There are 43 participants actually did not have depression and were not predicted to have depression; 7 participants actually did not have depression but were predicted to have depression; 49 participants actually have depression and were predicted to have depression; and 1 patient actually have depression but were predicted to not have depression.

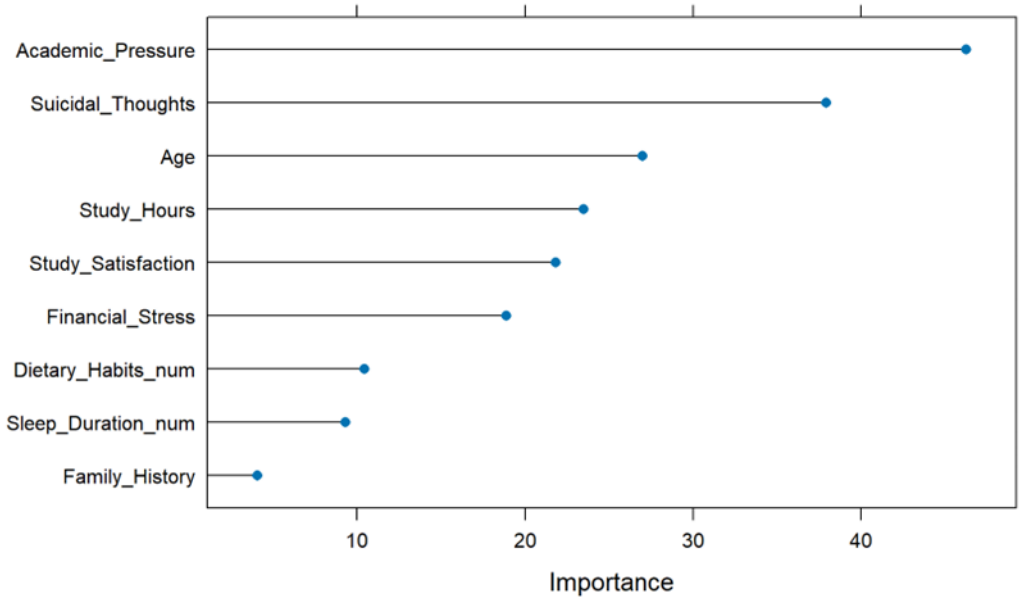


**Fig. 7** Variable importance of random forest model

Figure 7 illustrates the importance scores derived from the Random Forest model, ranking variables in descending order. The numerical values corresponding to these variables are 46.2517, 37.9148, 26.9994, 23.4851, 21.8231, 18.8802, 10.4468, 9.2929, and 4.0565 respectively. By comparing with logistic model's coefficients, variable such as Academic Pressure and Suicidal Thoughts again present a crucial role in the prediction of depression.

### 3.4. Result Comparison

Table 2 demonstrates a comparison of two models. The logistic regression model overtakes the random forest model in all parameters tested, with an accuracy of 0.9800, sensitivity of 1.0000, specificity of 0.9600, and AUC of 1.0000. In contrast, the random forest model achieves 0.9200 accuracy, 0.9800 sensitivity, 0.8600 specificity, and 0.9854 AUC. Based on the selected dataset, the findings suggest that logistic regression detects student depression more reliably and with fewer false positives, which is critical for early intervention. As a result, this comparison research emphasizes the effectiveness of logistic regression in identifying at-risk students while also confirming the potential utility of random forest as a competitive alternative in educational mental health evaluation.

**Table 2**. Two models comparing

| Model | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|
| Logistic Regression | 0.9800 | 1.0000 | 0.9600 | 1.0000 |
| Random Forest | 0.9200 | 0.9800 | 0.8600 | 0.9854 |

## 4. Conclusion

In conclusion, by comparing the performance of two models, Logistic Regression using elastic net presented a better prediction on test set, thereby may helping educational institutions and mental health professionals in early intervention and support strategies for students. The results of this study show that machine learning techniques have significant promise for predicting student depression by identifying crucial factors from a large dataset of demographic, academic, and lifestyle variables. The study, which used a sample of 502 people, highlights the importance of factors such as academic pressure, suicidal ideation, and age in shaping mental health outcomes. This approach provides a strong foundation for converting raw data into actionable insights, allowing the development of evidence-based solutions that can improve mental health support systems. This approach provides a solid framework for converting raw data into usable insights, allowing for the development of

evidence-based plans to improve mental health support systems for students. Despite the promising findings, the study admits inherent limitations due to the sample size and scope of factors evaluated, implying that more research with more diversified predictors and larger datasets is needed. Furthermore, investigating new modeling techniques or ensemble methods may improve prediction accuracy and robustness. Overall, this study adds to the literature on the use of data-driven methodologies in mental health diagnostics, highlighting the importance of quantitative analysis in addressing the widespread problem of student depression and informing future educational policies and practices.

## References

[1] Bernal-Morales B, Rodríguez-Landa J F, Pulido-Criollo F. Impact of anxiety and depression symptoms on scholar performance in high school and university students. In A fresh look at anxiety disorders. IntechOpen, 2015.

[2] Ibrahim A K, Kelly S J, Adams C E, et al. A systematic review of studies of depression prevalence in university students. Journal of psychiatric research, 2013, 47(3): 391-400.

[3] Deng Y, Cherian J, Khan N U N, et al. Family and academic stress and their impact on students' depression level and academic performance. Frontiers in psychiatry, 2022, 13: 869337.

[4] Mackenzie S, Wiegel J R, Mundt M, et al. Depression and suicide ideation among students accessing campus health care. American journal of orthopsychiatry, 2011, 81(1): 101.

[5] Pascoe M C, Hetrick S E, Parker A G. The impact of stress on students in secondary school and higher education. International journal of adolescence and youth, 2020, 25(1): 104-112.

[6] Shankar N L, Park C L. Effects of stress on students' physical and mental health and academic success. International Journal of School & Educational Psychology, 2016, 4(1): 5-9.

[7] McCloud T, Bann D. Financial stress and mental health among higher education students in the UK up to 2018: rapid review of evidence. J Epidemiol Community Health, 2019, 73(10): 977-984.

[8] Kempfer S S, Fernandes G C M, Reisdorfer E, et al. Epidemiology of depression in low income and low education adolescents: a systematic review and meta-analysis. Grant Med J, 2017, 2(04): 067-077.

[9] Li T M H, Li C T, Wong P W C, et al. Withdrawal behaviors and mental health among college students. Psicol Conductual, 2017, 25(1): 99-109.

[10] Roberts R E, Duong H T. The prospective association between sleep deprivation and depression among adolescents. Sleep, 2014, 37(2): 239-244.

[11] Ljungberg T, Bondza E, Lethin C. Evidence of the importance of dietary habits regarding depressive symptoms and depression. International journal of environmental research and public health, 2020, 17(5): 1616.

[12] Milne B J, Caspi A, Harrington H L, et al. Predictive value of family history on severity of illness: the case for depression, anxiety, alcohol dependence, and drug dependence. Archives of general psychiatry, 2009, 66(7): 738-747.