# Predicting Depression Risk Using Logistic Regression: A Case Study Based on Academic and Lifestyle Factors

Xianwen Jiang *

Department of Animation, Tokyo Polytechnic University, 2-9-5 Honcho, Nakano, Tokyo, Japan

* Corresponding Author Email: A2225045@st.t-kougei.ac.jp

**Abstract.** Depression has increasingly become a serious public health issue, with students and working professionals often among the most affected. Academic demands and financial challenges have been identified as key stressors, and recent global surveys show a noticeable rise in depression rates among young people, partly due to heavier workloads and reduced social support. In this study, logistic regression is applied to perform binary classification of depressive tendencies using the publicly available "Playground Series S4E11" dataset from Kaggle. This dataset contains demographic details along with academic and psychological indicators. The analysis followed a structured process—covering data cleaning, feature transformation, and training of the logistic regression model. Evaluation of the model using accuracy, recall, and Receiver Operating Characteristic-Area Under Curve (ROC-AUC) produced scores of 84%, 88%, and 0.913, respectively. Among the input variables, suicidal thoughts, academic or work pressure, and financial stress were identified as the most influential indicators. These findings indicate that even relatively simple and transparent models can serve as effective tools for the early identification of depression. When incorporated into digital systems, such approaches could help deliver timely support within educational institutions and healthcare environments.

**Keywords:** Depression Prediction, Logistic Regression, Mental Health Screening, Academic and Lifestyle Factors, Machine Learning.

## 1. Introduction

Hundreds of millions of people beat with the world, sufferers go unnoticed, untreated because still so much social stigma and since regular check-ups for mental health are not kept. More than 280 million people around the world presently suffer from depression — a leading cause of disability, according to the World Health Organization [1]. In academic and professional settings, pressures such as demanding workloads, financial challenges, and insufficient social support often intensify the risk of psychological decline.

In recent years, data-driven approaches have increasingly been explored within mental health research, showing promising outcomes. Among these, logistic regression stands out as both interpretable and dependable for binary classification, which explains its frequent use in medical and psychological studies. Recent research has illustrated its utility in detecting depression from structured data. For example, Qiang et al. demonstrated that explainable machine learning techniques—including logistic regression—can effectively identify key risk factors and anticipate mood changes in college students [2]. Similarly, Mimikou et al. reported the value of explainable models in predicting depression [3]. Mamun et al. applied predictive modeling that incorporated geographic and behavioral data to identify depression and anxiety trends among university students, highlighting the diversity of potential input variables [4].

Beyond the student context, Massell et al. used machine learning to forecast the initial onset of major depressive disorder in middle-aged adults [5], while Cruz et al. adopted a Naive Bayes–based framework to classify student mental health status [6]. In addition, Yang et al. demonstrated that explainable AI techniques can be applied to time-series mental health data to maintain predictive accuracy while ensuring interpretability—an essential factor in clinical applications [7].

This study uses a logistic regression approach on an open dataset that comprises academic and lifestyle features with the main aims of identifying which features are more strongly associated with risk of depression and performance evaluation using popular classification metrics. Results will

advance knowledge on how interpretable models can be leveraged for early detection and timely intervention, and later, in the long-term future, such methodologies might be merged into university counseling systems or mobile health applications for attaining scalable and approachable mental health surveillance.

## 2. Methodology

### 2.1. Dataset and Preprocessing

This study used the publicly available dataset from Kaggle's "Playground Series S4E11," which consists of over 140,000 individual responses. During data preprocessing, binary categorical features such as suicidal_thoughts and family_history were encoded into 0/1 format; pressure-related variables were merged into a single feature named academic_work_pressure; and satisfaction-related features were combined into a composite feature called study_job_satisfaction. Missing values in the Study Satisfaction column were imputed using mode, and all features were standardized using the StandardScaler. This preprocessing ensured that the model could learn from balanced and normalized inputs while preserving the interpretability of individual features.

### 2.2. Model Implementation

Implemented a logistic regression classifier using the liblinear solver with C=0.5 and max_iter=1000. The target variable is binary (0: not depressed, 1: depressed). The dataset was split into training (80%) and testing (20%) sets using stratified sampling to maintain class balance. Model training and evaluation were conducted using Scikit-learn, a well-established Python library for machine learning. Alternative models such as Naive Bayes have also shown promising results for depression status classification among students [6], although logistic regression remains more interpretable in feature-weighted contexts.

### 2.3. Evaluation Metrics

The model's performance was evaluated using several standard classification metrics. Accuracy reflects the overall proportion of correct predictions among all cases, while precision measures the proportion of predicted positive instances that were actually correct. Recall, which is particularly critical in mental health prediction, indicates the model's ability to correctly identify individuals truly at risk of depression. The F1-score serves as the harmonic means of precision and recall, providing a balanced assessment of the model's classification performance. The ROC-AUC score summarizes the model's discriminatory ability between the two classes across various thresholds, with values closer to 1 indicating superior performance. Additionally, the confusion matrix was analyzed to visualize the distribution of true positives, false positives, true negatives, and false negatives. Among these, recall holds special importance, as failing to identify someone at risk (i.e., producing a false negative) could have serious consequences in real-world screening scenarios.

### 2.4. Feature Importance

This study analyzed the logistic regression coefficients to interpret the most influential features contributing to predictions. Positive coefficients indicate increased likelihood of depression, while negative coefficients may reflect protective factors. Similar patterns of feature significance have been reported in clinical studies using logistic regression to identify psychological risk in cardiovascular patients [8].

# 3. Results

## 3.1. Classification Performance

The logistic regression model demonstrated robust classification capabilities on the test set. As summarized in Table 1, the model achieved a precision of 0.85 and recall of 0.88 for the depressed class (Class 1), indicating high sensitivity in detecting individuals at risk. The weighted average F1-score was 0.84, suggesting a balanced trade-off between precision and recall. These metrics reflect the model's practical utility in early depression screening, where minimizing false negatives is critical.

**Table 1.** Classification Metrics by Class.

| Metric | Class 0 (Not Depressed) | Class 1 (Depressed) | Weighted Avg |
|---|---|---|---|
| Precision | 0.82 | 0.85 | 0.84 |
| Recall | 0.79 | 0.88 | 0.84 |
| F1-score | 0.81 | 0.87 | 0.84 |
| Support | 2312 | 3267 | 5579 |

The logistic regression model produced the following results on the test set: these results indicate strong performance, particularly in identifying individuals at risk of depression. The high recall score ensures that the model captures most cases of depression, minimizing the chances of false negatives.

## 3.2. Confusion Matrix

As shown in Table 2, the confusion matrix illustrates that the model correctly identified 2878 cases of depression and 1822 non-depressed cases. A total of 389 false negatives and 490 false positives were observed. The relatively low number of false negatives is crucial in mental health applications, as undetected cases of depression can delay timely intervention and support.

**Table 2.** Confusion Matrix Results.

| | Predicted: Not Depressed | Predicted: Depressed |
|---|---|---|
| Actual: Not Depressed | 1822 | 490 |
| Actual: Depressed | 389 | 2878 |

The confusion matrix demonstrates a relatively low number of false negatives (389), which is especially important in clinical or educational screening scenarios where missing a true depressive case can lead to harmful consequences.

## 3.3. Experimental Results

The top eight most influential features, ranked by the absolute value of their logistic regression coefficients, are shown in Table 3. Features such as suicidal thoughts and academic/work pressure show strong positive associations with depression. On the other hand, factors like study/job satisfaction and age had negative coefficients, suggesting a potential protective effect.

**Table 3.** Top Eight Logistic Regression Coefficients.

| Feature | Coefficient |
|---|---|
| suicidal_thoughts | 1.204 |
| academic_work_pressure | 1.158 |
| financial_stress | 0.801 |
| study_job_satisfaction | -0.339 |
| work/study_hours | 0.431 |
| age | -0.533 |
| CGPA | 0.114 |
| family_history | 0.118 |

This analysis confirms that psychological stressors and workload significantly influence depression risk. Interestingly, a higher CGPA had a small but positive association, suggesting that academic performance pressure may contribute to stress, even when performance is objectively high.

## 4. Discussion

Despite its strengths, the model faces several limitations. First, the dataset relies on self-reported responses, which may introduce bias or inaccuracies due to stigma, misunderstanding, or personal interpretation of survey questions. Second, the dataset does not include certain variables that are commonly associated with mental health, such as sleep quality, peer support, personality traits, and history of trauma. Third, the model's binary classification simplifies depression into a yes/no outcome, whereas mental health typically exists along a spectrum. More nuanced models could better reflect real-world clinical conditions. Lastly, the generalizability of the model is limited by its single-source dataset. Without validation on external or cross cultural populations, the robustness of the model remains uncertain.

To address these issues, future research should consider integrating multi-modal data such as wearable sensor outputs, online behavior, or medical history. Although logistic regression is highly interpretable, more sophisticated models such as ensemble learning or deep neural networks may achieve higher predictive accuracy. Tools like SHAP and LIME can help explain predictions from these complex models [9, 10]. The effectiveness of such models during crisis events, such as during COVID-19 lockdowns, has already been demonstrated in prior student-focused studies [11]. Ethical considerations, especially regarding fairness, transparency, and privacy must also be addressed in the deployment of AI-based mental health screening tools.

## 5. Conclusion

This study demonstrates the effectiveness of logistic regression in predicting depression risk using academic and lifestyle-related survey data. With a ROC-AUC of 0.913 and high recall for the depressed class, the model provides a practical solution for early screening. The interpretability of logistic regression is a significant advantage in mental health contexts, allowing stakeholders to understand risk factors clearly.

Future studies should explore more complex models and broader datasets while maintaining explainability. Techniques such as SHAP or LIME can assist in model interpretation even with deep learning. Integrating richer feature sets including behavioral, physiological, and social data could further improve accuracy. Ethical considerations must be addressed when deploying such models in sensitive domains.

The current model serves as a strong baseline for depression risk screening and could be readily deployed in school counseling platforms or workplace wellness systems to identify individuals who may benefit from early support.

## References


[1] World Health Organization: "Depression: Key facts", https://www.who.int/news-room/fact-sheets/detail/depression, accessed 6 August 2025.

[2] Q. Qiang, J. Hu, X. Chen, et al. "Identifying risk factors for depression and positive/negative mood changes in college students using machine learning," *Front. Public Health*, 2025, 13, 1606947.

[3] C. Mimikou, C. Kokkotis, D. Tsiptsios, et al. "Explainable Machine Learning in the Prediction of Depression," *Diagnostics (Basel)*, 2025, 15, (11), 1412.

[4] M. A. Mamun, F. Al-Mamun, M. E. Hasan, et al. "Predictive Modeling of Comorbid Depression and Anxiety Symptoms Among Prospective University Students: A GIS-Based and Machine Learning Study," *Psychol. Rep.*, 2025, Advance online publication.

[5] J. Massell, M. Preisig, M. Miché, et al. “Prospective prediction of first onset of major depressive disorder in midlife using machine learning”, *Soc. Psychiatry Psychiatr. Epidemiol.*, 2025, Advance online publication.

[6] F. T. Cruz, E. E. Flores, S. J. Quispe, “Prediction of depression status in college students using a Naive Bayes classifier based machine learning model,” *Proc. Int. Conf. on Emerging Trends in* Computing *and Information Technology*, Lima, Peru, March 2023, pp. 87–92

[7] Y. Yang, L. Wen, L. Li, “Explainable AI for time series prediction in economic mental health analysis,” *Front. Med.*, 2025, 12, 1591793.

[8] L. Meng, Q. Zhu, F. Ma, et al. “Logistic regression analysis of risk factors for anxiety and depression in patients with coronary heart disease and subclinical hypothyroidism,” *Sci. Rep.*, 2024, 14, (1), pp. 27781.

[9] P. Prakrankamanant, S. Watanabe, E. Chuang, “Explainable Depression Detection using Masked Hard Instance Mining,” *arXiv preprint*, 2025, abs/2505.24609

[10] E. Kerz, S. Zanwar, Y. Qiao, et al. “Toward explainable AI (XAI) for mental health detection based on language behavior,” *Front. Psychiatry*, 2023, 14, 1219479.

[11] L. C. López Steinmetz, M. Sison, R. Zhumagambetov, et al. “Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine,” *Front.* Psychiatry, 2024, 15, 1376784.