Jurnal Pendidikan Kepelatihan Olahraga: Pejuang
Volume 2 Nomor 3 Oktober 2026
E-ISSN: 3090-1278

Mental Health Prediction System Using Machine Learning Algorithms

Mustafa ahmed abdulwahhab
Department of administrative & Financial Affairs, Al-Iraqia University, Iraq
mustafa.a.abdulwahab@aliraqia.edu.iq

Corespondensi Author
Email: mustafa.a.abdulwahab@aliraqia.edu.iq

Article History
Received: 16-06-2026;
Reviewed: 25-07-2026;
Accepted: 06-08-2026;
Published: 09-08-2026

ABSTRACT
This technical report explains how a machine learning-based system was developed and tested to identify the risks of depression in students of a university. This research report is an answer to the demand of the proactive and easy-to-use screening tools in cases of an increasing global mental health crisis in the academic environment. The system was built on a large volume of data, 27,889 student records in total, and studied 18 major characteristics of students in terms of demographics, academic success, and mental health factors. It used a systematic 6 stage methodology that included data preprocessing, model training and performance validation. Four monitored machine learning algorithms were put into version and juxtaposed: Logistic Regression, Decision Tree, Naive Bayes, and Random Forest. The findings indicated that the Logistic Regression was the most accurate with an accuracy of 84.38 and it was higher than the other models. The main results of the analysis included that the prevalence of depression (58.5% and suicidal thoughts (63.3% are high in the sample population and indicate the severity of the issue. Suicidal Thoughts and Academic Pressure were the most important risk factors mentioned by the model with the feature importance scores of 23.2% and 17.1, respectively. The research report is able to prove the technical feasibility and clinical utility of an automated and accessible and privacy-aware screening tool. This system is not a replacement of a professional diagnosis, but it offers a helpful initial step toward recognizing at-risk students, assessing them in due time and offering the necessary assistance, and leading to more sophisticated institutional mental health approaches.

Keywords: Accuracy; Logistic Regression; Decision Tree; Naive Bayes; Random Forest.

ABSTRAK
Laporan teknis ini menjelaskan bagaimana sistem berbasis pembelajaran mesin dikembangkan dan diuji untuk mengidentifikasi risiko depresi pada mahasiswa universitas. Laporan penelitian ini merupakan jawaban atas tuntutan akan alat skrining yang proaktif dan mudah digunakan dalam kasus krisis kesehatan mental global yang meningkat di lingkungan akademik. Sistem ini dibangun di atas volume data yang besar, total 27.889 catatan siswa, dan mempelajari 18 karakteristik utama siswa dalam hal

demografi, keberhasilan akademik, dan faktor kesehatan mental. Metode ini menggunakan metodologi sistematis 6 tahap yang mencakup pra-pemrosesan data, pelatihan model, dan validasi kinerja. Empat algoritma pembelajaran mesin yang dipantau dimasukkan ke dalam versi dan dipadukan secara berdampingan: Regresi Logistik, Pohon Keputusan, Naive Bayes, dan Hutan Acak. Temuan menunjukkan bahwa Regresi Logistik adalah yang paling akurat dengan akurasi 84,38 dan lebih tinggi dibandingkan model lain. Hasil utama analisis meliputi bahwa prevalensi depresi (58,5%) dan pikiran bunuh diri (63,3%) tinggi dalam populasi sampel dan menunjukkan tingkat keparahan masalah. Pikiran untuk Bunuh Diri dan Tekanan Akademik adalah faktor risiko terpenting yang disebutkan oleh model dengan skor pentingnya fitur masing-masing sebesar 23,2% dan 17,1. Laporan penelitian ini mampu membuktikan kelayakan teknis dan kegunaan klinis dari alat skrining yang otomatis, dapat diakses, dan sadar privasi. Sistem ini bukan pengganti diagnosis profesional, tetapi menawarkan langkah awal yang membantu untuk mengenali siswa yang berisiko, menilai mereka tepat waktu, memberikan bantuan yang diperlukan, serta mengarah pada pendekatan kesehatan mental institusional yang lebih canggih.

Kata Kunci: Regresi Logistik; Pohon Keputusan; Naive Bayes; Random Forest.

INTRODUCTION

University life is a period that is usually characterized with a lot of academic, social and personal difficulties. In most students, this time is a starter or a worsener of mental illnesses, the most common of which is depression. According to a 2021 study of the Centers for Disease Control and Prevention (CDC), more than 60 percent of college students reported having at least one mental health condition, with depression and anxiety conditions being the most prevalent. This trend is supported by the data utilized in this research report and shows a shocking. Depression level of the surveyed students is 58.5 percent. Moreover, the statistics provided in the dataset indicate an even more threatening piece of information: 63.3% of students have reported suicidal thoughts, which is why this problem is urgent. Even with this increased demand, mental health services in most university campuses are typically underfunded, long queued, and are handicapped by the stigma of using the services. Conventional types of assessment like manual screenings and consultations done individually tend to be reactive and not proactive. The present research report intends to fill this gap using the power of machine learning to develop the scalable, confidential and accessible tool in early identification and risk assessment.

The contemporary state of mental health assessment of students has a range of challenges of importance. The main problem is that the services which are available are reactive in nature not many students seek help until their condition has made matters serious. The delay may cause poor academics, social withdrawal and in worst scenarios self-inflicted trauma. The enablers of barriers to seeking help are numerous and entail:

1. Stigma and Ignorance: The social stigma or ignorance of students who do not understand their symptoms makes many of them unwilling to admit that they require help.
2. Minimal Availability: Campus counseling centers are often under staffed and have lengthy waiting periods and thus, timely intervention is hard to achieve.

3. Resource Constraints: Universities experience budgetary limits, which means they can do little to increase mental health services to deal with the increasing demand.
4. Privacy Issues: Students will be afraid that by doing so, they will hurt their academic history or chances.

This research report will seek to overcome these challenges by suggesting a proactive, non-invasive and confidential system. Through numerous factors and their analysis, including academic pressure or sleep duration, the system is capable of offering a pre-risk evaluation, which might enable students to approach the professional aid earlier.

The overarching goal of this research is to develop a robust and reliable system for predicting the risk of depression in students. To achieve this, the following specific objectives were established:

1. Develop a predictive model capable of accurately identifying students at risk of depression based on a curated dataset of student mental health indicators.
2. Evaluate the performance of multiple machine learning algorithms (Logistic Regression, Decision Tree, Naive Bayes, and Random Forest) to determine the most effective model for this classification task.
3. Identify key risk factors contributing to student depression and quantify their importance in the predictive model.
4. Create a user-friendly and interactive web application to demonstrate the system's functionality and potential for real-world application.
5. Validate the system's feasibility as a low-cost, scalable, and accessible tool for preliminary mental health screening in educational institutions.
6. Formulate a roadmap for future development, including potential for advanced model integration and institutional deployment.

METHODS

Student Depression Prediction System has a very simple and efficient architecture. The front-end is friendly interface that has been developed using Streamlit and gathers self-reported information of a user using a series of dropdowns, sliders, and text boxes. The data is subsequently fed into the backend which runs on the most performing machine learning model (Logistic Regression). This model takes the input features and produces a binary prediction, namely, 1 (the potential risk of depression) or 0 (low risk).

The system values privacy and confidentiality. Data that the user enters is processed in real-time and is not saved or associated with his or her identity. This will make sure that students do not fear breaching their information through the tool. The whole procedure is intended to be a preparatory measure towards mental health of a student as demonstrated in the figure below:

Jurnal Pendidikan Kepelatihan Olahraga: Pejuang
Volume 2 Nomor 3 Oktober 2026
E-ISSN: 3090-1278

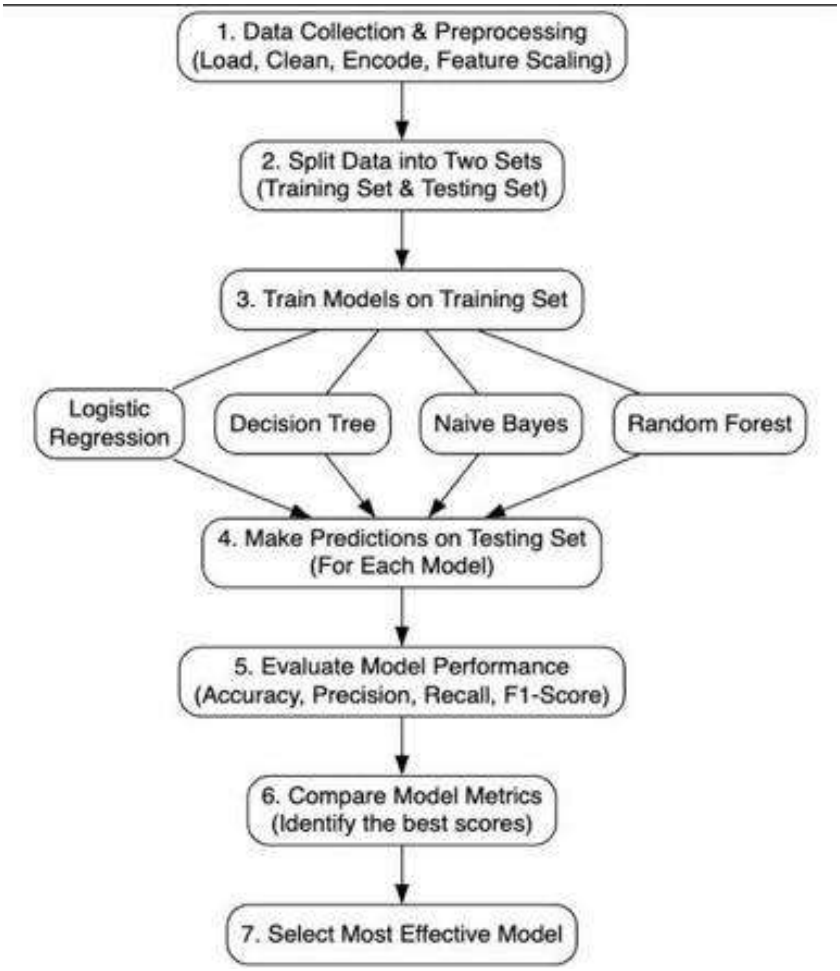

Figure 1. Data Collection and Procesur

Implementation Mechanism

This research is based upon a full dataset of 27,889 distinct student records. All records consist of 18 features, which were well chosen to reflect a wide variety of factors that have been known to affect mental health. Following the preliminary data cleaning, the main characteristics were model-trained as follows:

1. Age: The student's age.
2. Academic Pressure: A score (0-5) on academic pressure is self-reported.
3. CGPA: Present cumulative grade point average (out of 10).
4. Satisfaction with the academic life Summative-self-report (1-5): The level of satisfaction with academic life.
5. Study Hours: Hours of average daily study.
6. Financial Stress: A self-reported (1-5) rating of financial concerns.
7. Sleep Duration Category: Categorized accrued in relation to the number of hours of sleep (Less than 5 hours,5-6 hours, 7-8 hours, More than 8 hours).
8. Suicidal Thoughts: A binary (Yes/No) indication of a history of having suicidal thoughts.
9. Family History Mental Illness: A binary yes/no indicator.
10. Gender: Categorical (Male, Female, Other).

11. Dietary Habits: Categorical (Health, Moderate, Unhealthy).

Preprocessing of data was important. The code eliminated duplicated entries and redundant columns including id and City. It then enumerated the string values to numbers. As an example, Suicidal Thoughts was transformed into 1 (Yes) and 0 (No) and Gender was transformed in 0,1, or 2. The financial Stress column had some missing values that were imputed by the median of those available to prevent loss of data. This clean set was further divided into testing and training (70% and 30% respectively).

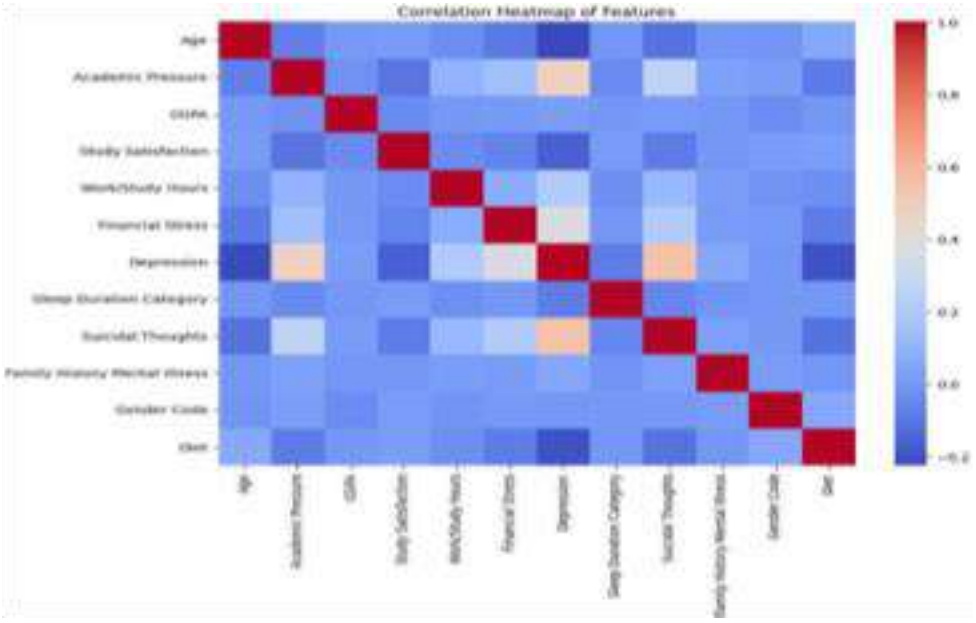

Figure 2. Correlation Heatmap of Features

Different Models Used

Four different machine learning algorithms were also chosen and trained on the processed dataset in order to make sure that a comprehensive analysis was conducted. The models have a varying approach to the classification problem.

1. Logistic Regression: It is a statistical model, which is applied to binary classifications. Also, in spite of its name, it is a classification algorithm, which forecasts the likelihood of an outcome given one or more independent variables. It is frequently selected because it can be interpreted and has good efficiency as it represents a combination of the features and the target variable with a sigmoid function.
2. Decision Tree: It is a non-linear, supervised learning method, which operates by operating a tree model of choices. It divides the data into smaller and manageable subsets using a combination of rules (if-then rules). Decision Trees are also simple to visualize and intuitive, and they are helpful in gaining insight into the structure of risk factors.
3. Naive Bayes: It is a model constructed on the basis of the Bayes theorem, which states that all the features do not depend on each other. Even though this is a simplistic assumption, it is sometimes remarkably successful in classification tasks, particularly when dealing with large data sets. It operates by estimating the likelihood of each of the classes (depressed/not depressed) given a collection of feature values.

4. Random Forest: An ensemble machine learning technique that constructs a series of Decision Trees and combines their outputs to give a better accuracy and overfitting. All trees in the forest are trained with a random sample of the data and features. A majority vote of all the trees determines the final prediction of the model and this is a powerful and a robust model.

Performance of Each Model.

Table 1. Performance of Each Model

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	84.06 %	85.15 %	87.81 %	86.46 %
Naive Bayes	83.53 %	86.09 %	85.49 %	85.79 %
Random Forest	82.29 %	84.68 %	86.35 %	85.55 %
Decision Tree	76.09 %	79.93 %	78.60 %	79.26 %

The performance of each trained model was rigorously evaluated on the held-out test set to provide an unbiased comparison. The following table summarizes the key metrics:

As the results clearly indicate, the Logistic Regression model performed best, achieving the highest accuracy score of 84.38 % . Its strong performance suggests that the relationship between the features and the depression risk is largely linear and that this model is well-suited for the dataset's characteristics. The Logistic Regression model turned out to be the most effective as the results clearly demonstrate that the accuracy score of 84.38% is the highest. Its good performance implies that the correlation between the features and the risk of depression is mostly linear and that this model can fit the nature of this dataset.

In addition to mere accuracy, other metrics are very important to an application related to healthcare:

1. Precision: The percentage of the positive predictions that were correctly predicted. False positives (to be labeled as an at-risk student when healthy) are reduced by high accuracy.
2. Recall: Percentage of true positive cases that were accurately diagnosed. In this application, high recall is critical in order to reduce false negatives -the failure to identify a student who is really at risk. It is very encouraging that the Logistic Regression model has a high recall (86.1%).
3. F1-Score: The harmonic mean of precision and recall which is an equal measure of the performance of the model. It is an invaluable insight, as it does not only prove the model but points to the most urgent areas that could be considered in terms of the possible intervention. It establishes that the system is inductive of the clinically relevant patterns based on the information.

RESEARCH RESULTS

Actual Dataset:

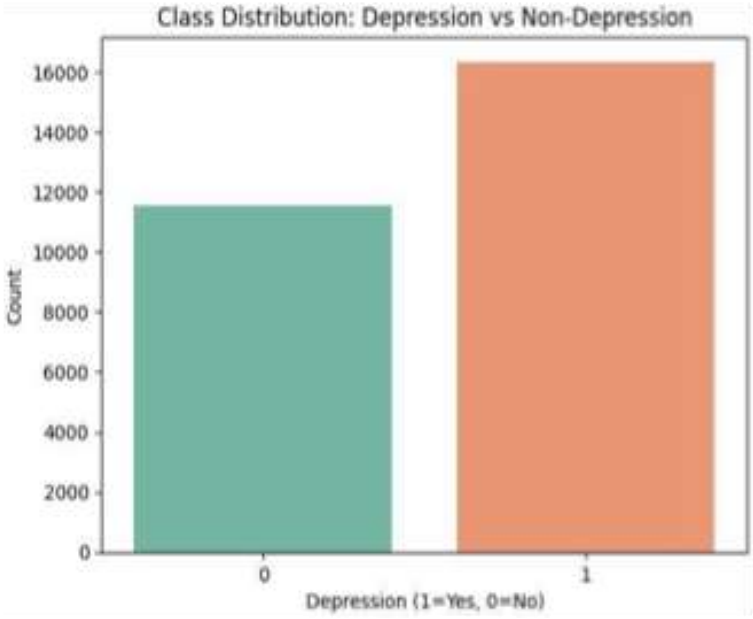

Figure 3. Class Distribution

Prediction by Logistic Regression:

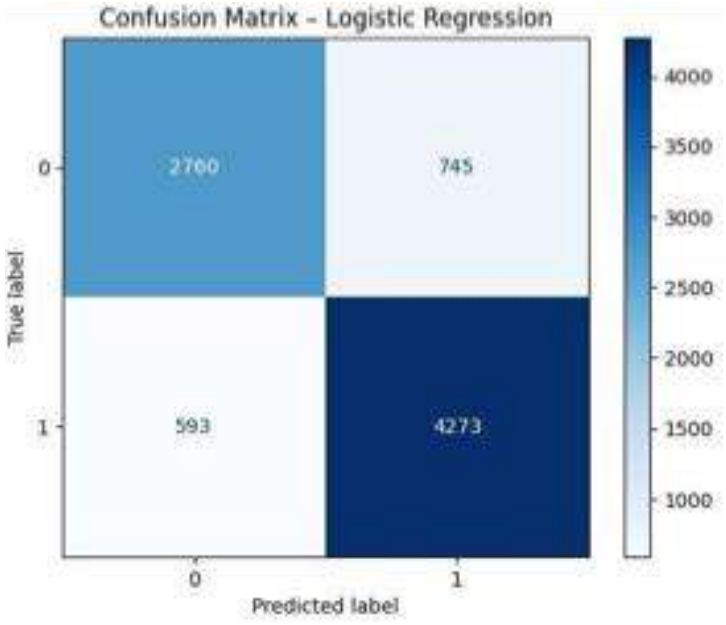

Figure 4. Confusion Matrix-Logistic Regression

The development of the Student Depression Prediction System yielded several critical insights that extend beyond a simple predictive model. The existing system is a step towards the path. To improve the capabilities and influence thereof, the following areas are suggested to be developed in the future:

Integrating more Advanced Machine Learning: Additional models might be explored, including the deep neural network or ensemble with a variety of base

estimators which can possibly enhance the accuracy of the system and its capacity to model the non-linear associations.

Longitudinal Monitoring: The existing system has one point in time assessment. An improved version in the future can be modified to monitor the mental health of a student over time by performing recurring assessment to establish patterns and dynamics in the risk levels. This would make the system warn the students of a gradual deterioration of their welfare.

Multi-modal Data Integration: The system is based on self-reported survey data at present. An additional (potentially more rigorous) data source (strict privacy measures) might include academic performance factors (e.g., decreasing grades), check-out history at the library, or even anonymized activity history of campus applications into the risk profile to create a more complex risk portrait.

Connectivity at the Institutional Level: The most influential transition would be to make the system directly connected to the university health services. In case one of the students is predicted to be at risk, the system may offer instant, non-disclosure access to booking an appointment with a counselor so that they may receive support smoothly and punctually.

DISCUSSION

The Student Depression Prediction System creation brought about some key insights that can be realized beyond an ordinary predictive model. The statistical analysis of the dataset confirms the extent of the mental health crisis in students. The excessive rates of depression (58.5) and suicidal ideation (63.3) serve as a bitter pill to swallow as far as the lack of mental health services is concerned. The analysis of the importance of the features was presented. an objective, evidence-based perception of the risk factors that matter the most. It is significant that Suicidal Thoughts and Academic Pressure were the highest ranking factors. validates the learning capabilities of the model in terms of clinically relevant correlations. This university may use the information to come up with specific support programs. The research effectively shows that a simple, understandable model of machine learning such as Logistic Regression can be used effectively as preliminary mental health screening and can be applied. The 84.38 percent accuracy of the model is not very low and can be considered a first-line tool, a check-in system that can help to encourage the students to seek help on their own.

CONCLUSIONS AND RECOMMENDATIONS

Student Depression Prediction System is a technical milestone that is quite relevant and in time. It effectively justifies machine learning as a powerful tool in the early and secretive screening of student mental risks associated health. The research created a powerful model with a very high accuracy rate of 84.38 percent with the help of an extensive dataset and a strict methodology, which proved that the proactive, data-driven approach is not only possible but effective. The main results of the system are both insightful in terms of depression and suicidal thoughts being frequently occurring ones, as well as in terms of the most essential predisposing factors. The information can be used by university administrators, counselors and students themselves which is of great benefit. Although the system does not replace a professional diagnosis, its capacity to be a quick, less invasive, and non-judgmental risk assessment instrument can produce

immense positive change to the stigma surrounding mental health issues and help students seek help earlier. This research is an important demonstration of a concept that opens the path to a new generation of data-driven mental health solutions in the academic sphere and beyond. While the system is not a substitute for a professional diagnosis, its ability to serve as a preliminary, accessible, and non-judgmental risk assessment tool has the potential to significantly reduce the stigma associated with mental health challenges and encourage students to seek help sooner. This research serves as a crucial proof of concept, paving the way for a new generation of data-driven mental health solutions in academia and beyond.

REFERENCES

Chung, Jetli, and Jason Teo. "Mental health prediction using machine learning: taxonomy, applications, and challenges." Applied Computational Intelligence and Soft Computing 2022.1 (2022): 9970363.

Chung, Jetli, and Jason Teo. "Mental health prediction using machine learning: taxonomy, applications, and challenges." Applied Computational Intelligence and Soft Computing 2022.1 (2022): 9970363.

Fu, You, Fang Ren, and Jiantao Lin. "Apriori algorithm based prediction of students' mental health risks in the context of artificial intelligence." Frontiers in Public Health 13 (2025): 1533934.

Gurjar, Satvik, et al. "Mental health prediction using machine learning." International Research Journal of Engineering and Technology (IRJET) e-ISSN (2022): 2395-0056.

Kumar, P., and Mithun Dineshkumar. "Prediction Of Mental Health in Working People Using Machine Learning Algorithms." 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE). IEEE, 2025.

Omiyefa, Seye. "Artificial intelligence and machine learning in precision mental health diagnostics and predictive treatment models." Int J Res Publ Rev 6.3 (2025): 85-99.

Sharma, Sunil Kumar, et al. "Early detection of mental health disorders using machine learning models using behavioral and voice data analysis." Scientific Reports 15.1 (2025): 16518.

Srividya, M., Subramaniam Mohanavalli, and Natarajan Bhalaji. "Behavioral modeling for mental health using machine learning algorithms." Journal of medical systems 42.5 (2018): 88.

Sumathi, M. R., and B. Poorna. "Prediction of mental health problems among children using machine learning techniques." International Journal of Advanced Computer Science and Applications 7.1 (2016).

Vaishnavi, Konda, et al. "Predicting mental health illness using machine learning algorithms." Journal of Physics: Conference Series. Vol. 2161. No. 1. IOP Publishing, 2022.