Student Depression Prediction Using Feature Selection and Logistic Regression Algorithm

Dileep Kumar Sharma[1], Aaryaa Khatiwada[2], Tanisha Thapa[3], Manaswee Niraula[4], Abhilasha Sharma[5]

{dileep.kumar@chitkarauniversity.edu.in[1], aarya1612.be23@chitkarauniversity.edu.in[2] , tanisha1620.be23@chitkarauniversity.edu.in[3] , manaswee1613.be23@chitkarauniversity.edu.in[4] , abhilasha.sharma@chitkarauniversity.edu.in[5]}

Chitkara University School of Engineering and Technology[1,2,3,4,5], Chitkara University[1,2,3,4,5] Baddi, HP, India

Abstract. This study applies logistic regression to predict depression among students in various classes, including 12th grade, undergraduate, postgraduate, and PhD courses, within a specific age group. Initially, a dataset consisting of 27901 records, having 17 input features and a target feature, is acquired. Different techniques such as missing value handling, level encoding, standardisation and dataset balancing using random oversampling are employed to preprocess the dataset. In the next step to minimise the problem of overmodelling and computational complexity, 13 important features are selected using random forest as a feature selector. Hyperparameter tuning using grid search has been performed to select the regularisation parameter of the regression model. The selected features are used to train the logistic regression classifier. The efficacy of the algorithm is tested with four different models. It is seen that the suggested method excels other classifiers in relation to recommended parameters to classify the scholars under depression.

Keywords: Depression prediction, logistic regression, data pre-processing, machine learning, feature selection.

1 Introduction

Depression is becoming a serious health issue among students as they face pressure, isolation, and performance demands in academics from their parents and society. Research findings indicate that such students are mainly susceptible to mental health challenges, and depression remains one of the most reported problems [1], [2]. Conventional methods such as clinical interviews and surveys are helpful in identifying the students at risk; however, these methods are time-consuming, subjective and difficult to apply to a large scale population [3].

With the rapid increase in machine learning, there is tremendous scope for finding new ways to detect and respond to mental health issues and disorders. Supervised learning models, in particular, are showing promise in identifying students who may be struggling with depression, using data related to their academic performance, habits, and lifestyle choices [4] [5]. These models can pick on subtle patterns in data that may go unnoticed by the human eye, allowing

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362524

for earlier and more targeted interventions.
In this research, the authors have applied logistic regression (LR) in the classification of anxiety among the population of varying age groups. Feature engineering is done to find out the predominant features and assess which features affect depression the most [6]. The study is compared with four ML algorithms, namely random forest (RF) [7], support vector machines (SVM) [8], gradient boosting [9] and decision tree (DT) [10] using four ML classification performance parameters. To the best of the authors the combination of missing value handling, level encoding, standardisation, random oversampling and feature selection using random forest and hyperparameter tuning for the proposed method and dataset is not reported in the literature. Hence, the present study is an attempt to fill this gap.

2 Literature review

This section gives a short review of the conventional methods for the prediction of depression, as follows:
Cruz *et* al**.** [11] applied a Naive Bayes (NB) based classifier in the prediction of the level of depression among 519 university students, which achieved an accuracy of 78.03%. Lyu [12] applied a random forest model to find moderate and severe depression patients and used 23 features in the experiment. The model achieved a perfection of 86.8% and was better in depression prediction compared to SVM, NB, RF, and artificial neural network. It was observed that age, level of education, and living expenses influenced depression. Norouzi and Machado [13] used an SVM model to predict depression, anxiety, and stress. They achieved an accuracy of 99% for their datasets. Du *et* al. applied convolutional neural networks (CNN) to determine the anxiety level of college students using electroencephalogram data. The model classified the mental level as positive, negative, or neutral with a correctness of 97.54% and a harmonic mean of 98.35%, showing the superiority in predicting the mental health of scholars engaging EEG data [14].
In their comparison of SVM, LR, and other models for depression prediction, Sabouri et al. [15] discovered that SVM and LR had the best accuracy, at 83.32%. They concluded that basic machine learning models might assist physicians in identifying depression in its early stages. Zhong and Yadav [16] predicted depression in high school students in the United States using an ANN model. In contrast to the SVM model, which received a harmonic mean of 81.90%, the model obtained the highest harmonic mean score of 82.90%. Using a survey-based educational dataset, Muzumdar et al. [17] evaluated several machine learning models, including Naive Bayes, SVM, K-nearest neighbour (KNN), LR, stochastic gradient descent (SGD), DT, RF, extreme gradient boosting (XGBoost), and NGBoost. With an accuracy of 99% across the four datasets examined, the XGBoost model demonstrated the best performance. Additionally, it outperformed other methods in performance scores and estimated run-time, demonstrating its efficacy and efficiency in categorising the mental health of students.
Iparraguirre *et* al. [18] applied the LR model for the detection of depression among 781 college students. He compared the LR model with KNN and DT and obtained better F1 score, recall, and accuracy for the LR model as compared to other models. The study also revealed that male scholars are twice as expected to be obese, 2 among 5 scholars suffer from gentle anxiety, and male scholars typically have a higher risk of being overweight than female scholars. Obese students are also 2.5 times more expected to suffer from anxiety. Singh *et* al. [19] used ML models like DT, RF, SVM, AdaBoost, NB, LR, and KNN to predict stress levels among 843 students. These students answered 28 questions checked by the experts. The SVM model

performed the best which achieved the highest accuracy of 95%, showing that it works well in detecting mental stress among students. Reece and Danforth [20] used machine learning techniques to study more than 43,000 Instagram photos from 166 people. They looked at things like colours in the photos, metadata, and facial features using algorithms. Their models, including logistic regression, were able to spot signs of depression with 70% accuracy which was better than the average success rate of general doctors, which is only about 42%.

3 Motivation

Traditional methods of diagnosis and monitoring, such as those based on clinical interviews and pen-and-paper questionnaires, are typically unable to extract deeper personality-based behavioural patterns that frequently anticipate early risk. Still, little progress has been made in applying psychological methods to practical and predictive applications. Data Scientists can work toward creating systems that can forecast a person's likelihood of experiencing mental health issues by combining clinical data with ML. This can offer supplementary resources that enable more active and accessible mental health care. This research is an attempt towards achieving this goal.

4 Research Contribution

This study explores how different stressors—academic pressure, work-related stress, lifestyle choices, financial struggles, and family history of mental illness—affect students' mental health. The research contributions are given as follows:

A logistic regression-based machine learning model is proposed, which uses a unique combination of data preprocessing, such as missing value handling, level encoding, standardisation, random oversampling, feature selection using random forest and hyperparameter tuning to effectively predict depression based on these stress factors.

An analysis of feature importance is done using random forest to determine which characteristics most contribute to model prediction.

5 Dataset details

The dataset was obtained from Kaggle [21] and contains 27901 records of the depression dataset. The dataset carries 17 input and 1 target attributes. The input features are: id, Have you ever had suicidal thoughts? Academic Pressure, Financial Stress, CGPA, Age, City, Work/Study Hours, Degree, Study Satisfaction, Dietary Habits, Sleep Duration, Family History of Mental Illness, Gender, Profession, Work Pressure, and Job Satisfaction. The target feature is Depression.

6 Methodology

Figure 1 represents diagrammatically the different stages employed in the classification of the depression dataset. The details of various steps are given as follows:

6.1 Step I: Data preprocessing

The dataset is first downloaded from Kaggle [21] and pre-processed. The student ID (id column of the dataset) and the profession column are removed as they do not contribute any information to the ML model. Many columns of the dataset, such as city, gender, sleep duration, degree, etc., contain categorical values; hence they are converted into numbers through label encoding. Then, any missing readings are replaced with the median term to avoid errors during training and to improve model performance. After that, the population is split into input matrix (X) and the depression column in the dataset is chosen as the target (y).

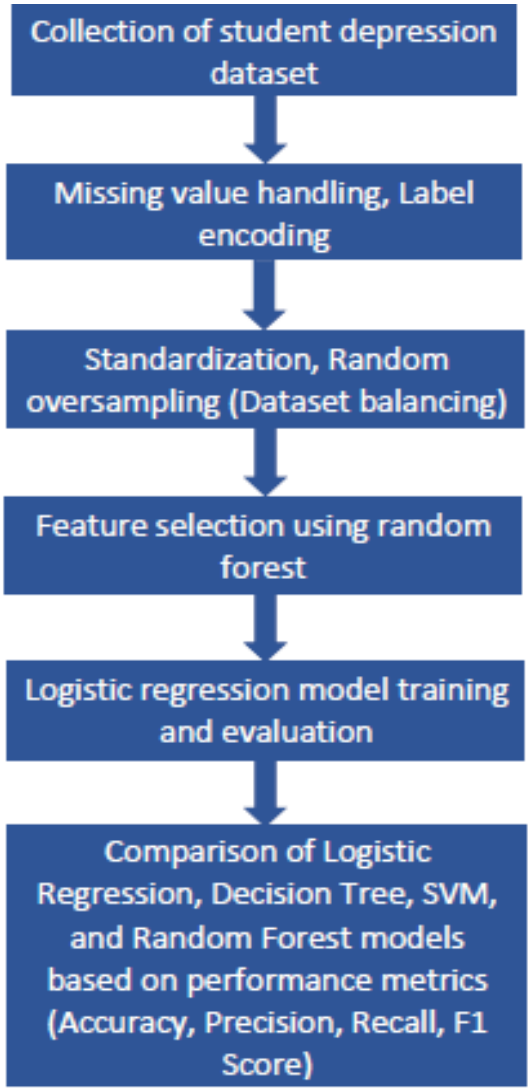


Fig.1. ML-based depression detection: Methodology

6.2 Step II: Split and Standardise

In this step, the population division ratio of 80:20 is utilised. Since some features, such as age, CGPA, study hours, and financial stress, are on different scales, they are standardised so that one feature does not dominate the others during model training. This helps enhance the model's performance. The data is standardised through the following formula in equation (1):

$$S = \frac{f_{value} - mean}{Standard\ deviation} \quad (1)$$

Where S is the standardised form and f_{value} is the attribute that is being standardised.

6.3 Step III: Balance the Dataset

The dataset is found to be imbalanced, so to fix this, the random oversampling process is used to duplicate minority class samples in the training set. This helps avoid bias toward the majority class during model training.

6.4 Step IV: Feature Selection with Random Forest

A random forest model is applied to evaluate feature importance. This reduces the overfitting problem and improves the model's generalisation. Attributes having a significance score > 0.01 are chosen for dimensionality reduction and to remove less informative features. Feature importance is plotted in Figure 2. This visualisation helps to understand which attributes most influence depression in students. The bar plot visualises how important each feature is, as determined by the Random Forest model.

As seen from this chart, the most critical indicator of depression is related to suicidal thoughts, followed by Academic Pressure, Financial Stress, CGPA and Age. Four features were excluded from the dataset to reduce complexity, as they play a minor role in the model's training

6.5 Step V: Train Logistic Regression With Selected Features and Hyperparameter Tuning

The LR algorithm employed in the present research is explained as follows:
It is a supervised ML algorithm that is used to analyse binary outcomes when multiple input variables are present. It evaluates the relationships between one or more input attributes and a dependent attribute with two possible categories, analysing all variables together to clarify their relationships and avoid unwanted hidden influences [22]. Although it is similar to multiple linear regression, instead of predicting a continuous outcome, it is used when the response variable has only two possible outcomes [18]. The formula of logistic regression is given in equation (2) and is explained as follows:

The model estimates the probability $P(Y = 1|X)$ given the input features represented by a matrix X.

$$P(Y = 1|X) = y_{hat} = \frac{1}{1+e^{-(W_0 + W_1 * X_1 + W_2 * X_2 + \cdots + W_n * X_n)}} \quad (2)$$

Where y_{hat} is the probability $P(Y = 1|X)$ that output $Y = 1$.
In a vectorised form, equation (2) can be written as equation (3).

$$y_{hat} = \frac{1}{1+e^{-(W^T * X)}} \tag{3}$$

Where $W = (W_0, W_1, W_2 \dots \dots \dots \dots \dots, W_n)^T$ is the vector of weights, including bias W_0.

$X = (1,\ X_1,\ X_2, \dots \dots \dots \dots \dots, X_n)^T$ is the input vector with bias term included.

The decision rule for the prediction of binary output is given in equation (4) as follows:
$Y = 1\ if\ y_{hat} \geq 0.5, else\ Y = 0$ (4)

The LR model is made to learn, using the important attributes identified in step IV. A grid search with cross-validation is performed to find the best values for regularisation strength.

7 Performance Evaluation Parameters

7.1 Confusion Matrix

It helps in defining the functioning of a classification model, especially in binary classification models, where possible outcomes are classified as positive and negative. It consists of metrics to define the classifier: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Further, several metrics are calculated with the help of these outcomes to evaluate the classifier's quality [19]. It gives an outline of TP, FP, TN and FN in tabular form as shown in Table 1.

Table 1: Confusion matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

7.2 Precision (P)

It is a type of performance metric for classification models that measures correct positive predictions. It is essential when false positives have a higher impact on the outcome than other outcomes [23]. It is obtained by dividing TP by the sum of TP and FP [24] and is given in equation (5).

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{5}$$

7.3 Accuracy

It is a performance metric that can be defined as the rate of correct classifications. It calculates the performance of a classifier by computing the correct predictions out of the total predictions. It can also be misleading in the case of data imbalance [25]. It is the fraction of true predictions

to the summation of all predictions in percentage [23] and is defined in equation (6).

$$A = \frac{TP+TN}{TP+FP+TN+FN} * 100 \quad (6)$$

7.4 Recall (R)

It is a representative parameter to measure the number of true positive results that were precisely forecasted by the model. A high recall value yields a better model, as it means more actual positive results are identified. It ensures that as many positive instances as possible are correctly identified; hence, it is essential when false negatives have a higher impact on the outcome than any other outcome [25]. It is obtained by dividing TP by the sum of TP and FN [26] as expressed in equation (7).

$$\mathrm{R} = \frac{TP}{TP+FN} \quad (7)$$

7.5 F1-score

It is a widely used representative parameter for classification jobs, which fuses precision and recall into a single metric by calculating their harmonic mean as shown by equation (8). The higher the F1-score closer to 1 better the model's predictions. It helps to balance model performance in both positive and negative classes. It is crucial in situations like an imbalanced population [27].
].

$$\mathrm{F1}_{\mathrm{score}} = \frac{2(P*R)}{P+R} \quad (8)$$

8. Results and discussions

After training, the execution efficacy of the model is tested using parameters as discussed in Section 7. The model superiority is also tested using 4 other ML models, as manifested in Table 2. As perceived from this table, the proposed method performs well with a precision of 87.31%, an accuracy of 83.52% and an F1 score of 85.50% as contrasted to other schemes. The confusion matrix of the test dataset is also plotted and is shown in Figure 3. As seen from this matrix, 1949 cases had no depression and were correctly classified as cases with no depression. As seen from this figure, 526 cases had depression but were falsely classified as cases having no depression.

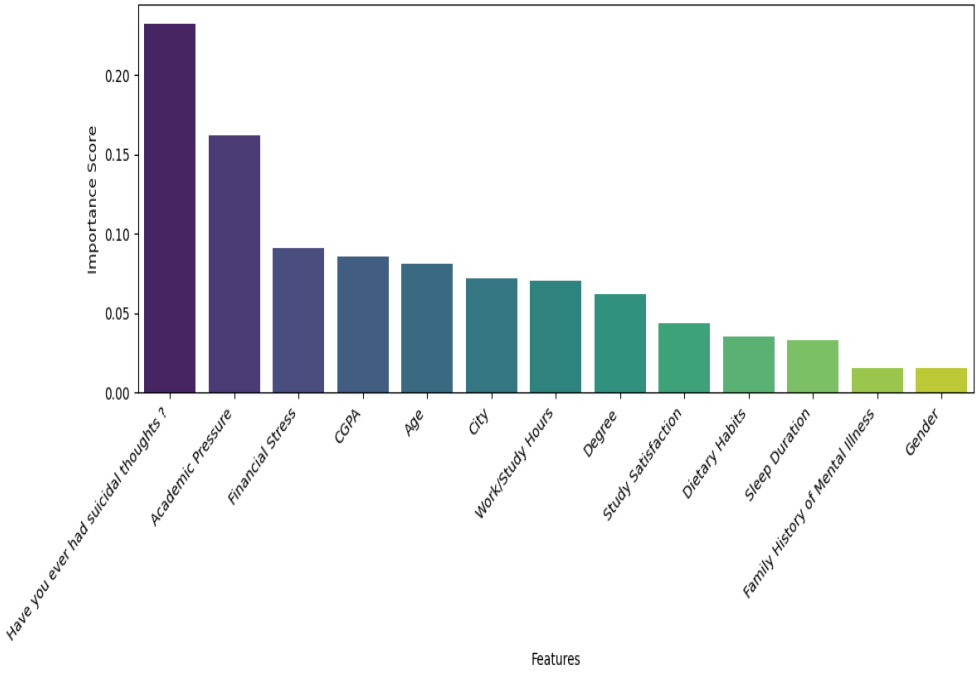


Fig. 2. Bar chart showing selected features using a random forest model

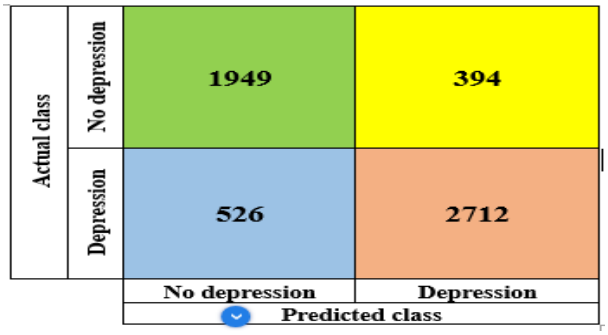


Fig. 3. Proposed method confusion matrix for the presented population

Table 2: Numerical evaluation of the suggested method with four conventional methods

Model	Accuracy	Precision	Recall	F1 Score
Support Vector Machine [8]	83.12%	86.65%	83.82%	85.21%
Decision Tree [10]	76.44%	80.23%	78.81%	79.51%
Random Forest [7]	82.87%	85.24%	**85.24%**	85.24%
Gradient boosting [9]	83.19	86.55	84.10	85.31
Proposed Logistic Regression-based model	**83.52%**	**87.31%**	83.76%	**85.50%**

Similarly, 394 cases had no depression but were falsely classified as the depression class. 2712 students were suffering from depression, and the model also predicted them as positive cases of depression.

9 Conclusion

This article proposed a logistic regression-based ML model for the prediction of depression among scholars. The dataset, having 17 input features and a target variable, is initially preprocessed using missing value handling, label encoding, standardisation and random oversampling. Features importance is evaluated using random forest. It is seen that suicidal thoughts and academic pressure lead to depression among students. Some of the input features were removed after computing feature importance. A grid search with cross-validation is performed to find the best value for the regularisation parameter. The LR model was practised using the preprocessed dataset. Performance is checked using precision, recall, accuracy and F1 score. It is seen that the proposed classifier outperforms in comparison to conventional classifiers and is better suited to predict depression among the students. The study offers critical perspectives for identifying depression at an early stage and developing effective intervention strategies, helping educational institutions and mental health professionals support student well-being.

10 Future Scope

In future clinical levels, such as formal stress test results or a psychologist's diagnosis, may be included in the population for accuracy improvement. Moreover, advanced deep learning networks or transformer-based models can be tested in subsequent studies.

References

[1] Association, A.C.H.: American college health association-national college health assessment II: Reference group executive summary spring 2019 American College Health Association, Silver Spring, MD (2019)
[2] Ibrahim, A.K., Kelly, S.J., Adams, C.E., Glazebrook, C.: A systematic review of studies of depression prevalence in university students. Journal of Psychiatric Research. 47(3), pp. 391–400 (2013)
[3] Kroenke, K., Spitzer, R.L., Williams, J.B.: The PHQ-9: validity of a brief depression severity measure. Journal of General Internal Medicine. 16(9), pp. 606–613 (2001)
[4] Al Hanai, T., Ghassemi, M.M., Glass, J.R.: Detecting depression with audio/text sequence modeling of interviews. pp. 1716–1720 (2018)
[5] Trotzek, M., Koitka, S., Friedrich, C.M.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering. 32(3), pp. 588–601 (2018)
[6] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (2017)
[7] Breiman, L.: Random forests. Machine Learning. 45(1), pp. 5–32 (2001)
[8] Vapnik, V.: Support-vector networks. Machine Learning. 20, pp. 273–297 (1995)
[9] Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of Statistics. pp. 1189–1232 (2001)
[10] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S.: Top 10 algorithms in data mining. Knowledge and Information Systems. 14(1), pp. 1–37 (2008)
[11] Cruz, F.T., Flores, E.E.C., Quispe, S.J.C.: Prediction of depression status in college students using a Naive Bayes classifier based machine learning model. arXiv:2307.14371. (2023)
[12] Lyu, H.: Application of machine learning on depression prediction and analysis. Applied Computational Engineering. 5, pp. 712–719 (2023)
[13] Norouzi, F., Machado, B.L.M.S.: Predicting mental health outcomes: A machine learning approach to depression, anxiety, and stress. International Journal of Applied Data Science in Engineering and Health. 1(2), pp. 98–104 (2024)
[14] Du, C., Liu, C., Balamurugan, P., Selvaraj, P.: Deep learning-based mental health monitoring scheme for college students using convolutional neural network. International Journal on Artificial Intelligence Tools. 30(06n08), pp. 2140014 (2021)
[15] Sabouri, Z., Gherabi, N., Nasri, M., Amnai, M., Massari, H.E., Moustati, I.: Prediction of depression via supervised learning models: performance comparison and analysis. International Journal of Online & Biomedical Engineering. 19(9). (2023)
[16] Zhong, N., Yadav, N.: What's Race Got to do with it? Predicting youth depression across racial groups using machine and deep learning. arXiv:2308.11591. (2023)
[17] Muzumdar, P., Basyal, G.P., Vyas, P.: An empirical comparison of machine learning models for student's mental health illness assessment. arXiv:2202.13495. (2022)
[18] Iparraguirre-Villanueva, O., Paulino-Moreno, C., Epifanía-Huerta, A., Torres-Ceclén, C.: Machine learning models to classify and predict depression in college students. International Journal of Interactive Mobile Technologies. 18(14). (2024)
[19] Singh, A., Singh, K., Kumar, A., Shrivastava, A., Kumar, S.: Machine learning algorithms for detecting mental stress in college students. IEEE. pp. 1–5 (2024)
[20] Reece, A.G., Danforth, C.M.: Instagram photos reveal predictive markers of depression. EPJ Data Science. 6(1), pp. 15 (2017)
[21] Kaggle: Student Depression Dataset. Available at: https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset?resource=download

[22] Sperandei, S.: Understanding logistic regression analysis. Biochemia Medica. 24(1), pp. 12–18 (2014)
[23] Vujović, Ž.: Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications. 12(6), pp. 599–606 (2021)
[24] Sharma, A., Pandit, S., Talluri, S.R.: Throughput prediction of fifth-generation cellular system using

hybrid feature selection and enhanced sequential decision tree machine learning algorithm. Wireless Networks. 31(3), pp. 3025–3042 (2025)
[25] Khakhar, P., Dubey, R.K.: The integrity of machine learning algorithms against software defect prediction. In: Artificial Intelligence and Machine Learning for Edge Computing. pp. 65–74 (2022)
[26] Sharma, A., Talluri, S.R., Pandit, S.: 5G Wireless Technology Throughput Prediction Using Ensemble Machine Learning Approach. Springer. pp. 72–81 (2024)
[27] Christen, P., Hand, D.J., Kirielle, N.: A review of the F-measure: its history, properties, criticism, and alternatives. ACM Computing Surveys. 56(3), pp. 1–24 (2023)