ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14737958

Predicting Student Depression Using Machine Learning

Piyush Agarwal[1]; Rahul S Mundaragi[2]; Rahul Sanjay Kohad[3]; Rithvik Allada[4]; Samarth R Bharadwaj[5]; and Dr. Shobha T[6]

[1,2,3,4,5,6]Department of Information Science and Engineering,
BMS College of Engineering, Bangalore, India

Publication Date: 2025/01/27

Abstract: The masses of information surrounding the consequent low academic performance and bad general health of students have brought the issue of depression into the limelight. Both academic stress and personal and social issues act as co-factors in the genesis of depression in students. However, there are a number of challenges with respect to identification of students who are at risk of developing depression owing to the sensitive nature of mental health issues and social stigma. This approach is employing state-of-the-art techniques to predict student depression by analyzing social engagement, academic, and lifestyle-related variables. Three ma- chine learning models were implicated in the study, Logistic Regression, Decision Tree Classifier and Random Forest. The data set consisted of demographic data, self-reported mental health assessments, and academic-related information. The outputs are passed to a single 'Ensemble' model to improve prediction accuracy. The purpose of this study is to develop a model that is accurate, reliable, and can timely detect student depression and offer useful information to teachers, psychologists, and state policies as a way of helping high-risk students timely. Hence, it aims towards a more positive academic environment.

Keywords*: Depression, Machine Learning, Ensemble Model, Academic Stress, Mental Health.*

How to Cite: Piyush Agarwal; Rahul S Mundaragi; Rahul Sanjay Kohad; Rithvik Allada; Samarth R Bharadwaj; Dr. Shobha T (2025). Predicting Student Depression Using Machine Learning. *International Journal of Innovative Science and Research Technology*, 10(1), 940-945. https://doi.org/10.5281/zenodo.14737958

I. INTRODUCTION

In the past years, one of the major issues that have come to life is mental health issues, which interferes with moral of people of society as well as people on the individual level. The causes of depression are multifaceted, involving genetic, biological, environmental, and psychological factors.[1] Particularly hoping to find solutions for the issue of depression, which is found to be most worrying problem , due to its ability to affect a student's study, social activity and even their general state of body. The proposed solution goes hand in glove with academic activities even if the joy is rewarding, the stress exposes the students to mental issues such as depression. To effectively tackle the problem, targeting students at risk and providing enough help is very important.

This study utilized machine learning methods to identify risk factors associated with depression and anxiety in school-aged children. The findings highlighted the significance of school violence, bullying, home violence, academic performance, and family income as key contributors to mental health issues.[2] Such complex issues such as depression can be predicted using structured data from students with the help of this program. This solution tries to find patterns in questionnaire responses about academic achievements, sociability and aspects of everyday life that may help in predicting depression.

The model integrates three machine literacy algorithms which are Multi Class Logistic Regression, Decision Tree, Random Forest Classifier. All these models are trained and estimated on pupil records dataset to see if a pupil is likely to suffer depression. To ameliorate the prophetic delicacy and trustability, the system integrates the workings of these models into one ensemble model that combines their capacities to produce stronger results.

The data employed in this exploration, which is appertained to as ' Student Depression Dataset', comprises of anonymized scholars' data and internal health analysis applicable features. These features include demographic, academic, cultures, and tone reported internal health pointers. This dataset forms the base for training, testing and validating the models. The core ideal of this program is to make a dependable system that can prognosticate pupil depression in an automated fashion. Such a system can be veritably useful for preceptors, counselors and decision makers in the early identification of the scholars who need

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14737958

support interventions. Also, this design demonstrates the effective uses of machine literacy for working global challenges, demonstrating its cross correctional uses.

Using data from yearly health assessments, this study created a machine learning model to forecast students' mental health problems. The study highlights how useful these models are for early detection and student mental health intervention techniques.[3]

II. PROBLEM STATEMENT

Depression among scholars is an overwhelming and growing concern affecting their academic, social, and personal lives. With growing competition and pressure, there is performance anxiety in academic institutions, thus causing emotional injuries and exposing them to depression of sorts. Every nation invests lot of money on education. However research survey on college students reports at any given time there will be 10 to 20% of student population suffering from psychological problems (Stress, Anxiety & Depression).[4] Some of the present methods of handling student depression tend to rely upon self- reporting or reporting by teachers and peers, and both methods sometimes can be inaccurate and tardy. Such types of styles may lose out on the details of progression and the risk factors associated with the development of depression, thus providing very little room for early intervention and, as a result, worsening student issues. What one needs urgently is an organized, data-backed approach that would highlight the early identification of potential students with undiagnosed depression.

The great challenge is that many colorful factors contributing towards the development of depression are closely interlinked: among them are academic performance, social conditioning, life actions, and demographic information. Assessing these factors must provide an appropriate scheme capable of handling huge datasets, correlating patterns, and making reliable predictions. This design hopes to fill this gap and develop a system based on machine learning that predicts depression among students. By using structured datasets and advanced algorithms, the system further aims to directly categorize students into either the depression-affected group or not at risk of being depressed.[5] Not only will the system help identify students at risk, but it will also show the significant factors causing the mental health problems.

The important questions which this design aims to address are: Would machine learning models effectively predict the depression in students on the basis of the inputs like the academic performance, life habits, and self-reported criteria? What factors contributed the most toward the probability of depression in students? Also, how can one take ad- vantage of the ensemble modeling techniques to improve the accuracy and reliability of predictions relative to a single machine learning model?

By answering these questions, this design aims at producing a practical outcome from preceptors-mentors, counselors, and policymakers for providing timely and effective support to scholars. The ultimate goal is the betterment of internal health problems of scholars in fostering a healthier and more supportive academic environment.[6]

III. LITERATURE REVIEW

Ensemble styles in machine input have shown great promise in enhancing the performance of model predictions in various fields, including internal health. Logistic Regression, Decision Tree Classifier, and Random Forest have arguably made rampant use of predictive analytics, but ensemble modeling provides a definite leap in accuracy and robustness among any grouped approaches. A case is presented by Alaimo et al. in which Decision Trees and Random Forest were an ensemble model using Decision Trees and Random Forest for predicting early signs of inner health conditions in scholars resulting in more robust and generalized prognoses when compared with single modeling steps. These styles in the field of student well-being and health proved particularly useful as they trained ensemble methods, such as boosted regression trees (GBM), to address the high variability in student behavior and external influences. Combining boosting methods such as graduate boosting with traditional decision tree classifiers enables ensemble methods to better handle imbalanced datasets. This combination exploits both the strengths of the classification model and the principle of iterative refinement and has been shown to reduce error ranges significantly in environments characterized as noisy and scarce (Smith et al.).

Prepare data correctly by normalizing and standardizing in order to produce coherent distributions that would allow for ensemble modeling outperformance. Johnson et al. in their paper detail some preprocessing techniques like data imputation and scaling that allowed the ensemble model to accurately predict depression levels of students when predicted based on academic and life factors.

IV. MODEL DESIGN

Three competing models will be put forth in this study: Random Forest, Logistic Regression and Decision Tree. Random Forest is an ensemble method whereby a number of individual decision trees are generally built from random subsets of training data. It makes predictions by combining the results of all trees which reduces bias and increases accuracy to avoid overfitting. It is also most useful in catching complex long-term relationships in the data. Next is Logistic Regression, a linear model that is often used for binary classification problems. By fitting the data to a logistic function, it produces the probability of an event occurring. Easy and intuitive to interpret; however, it assumes linear relationships between its features and output-target variable and this may affect its predictive performance on highly non-linear datasets. After this comes Decision Tree, a tree-like model in which branches are created based on feature values so that categories can be classified accordingly. It has the greatest intuitive space, effective visualization but succumbed to overfitting, especially in

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14737958

deep trees or noisy data. Thus, features of such models-that is the strength provided by Random Forest with a robustness, Logistic Regression in its simplicity, and interpretability provided by Decision Trees-could each be shared without partaking in the defect associated with either.

V. ARCHITECTURE

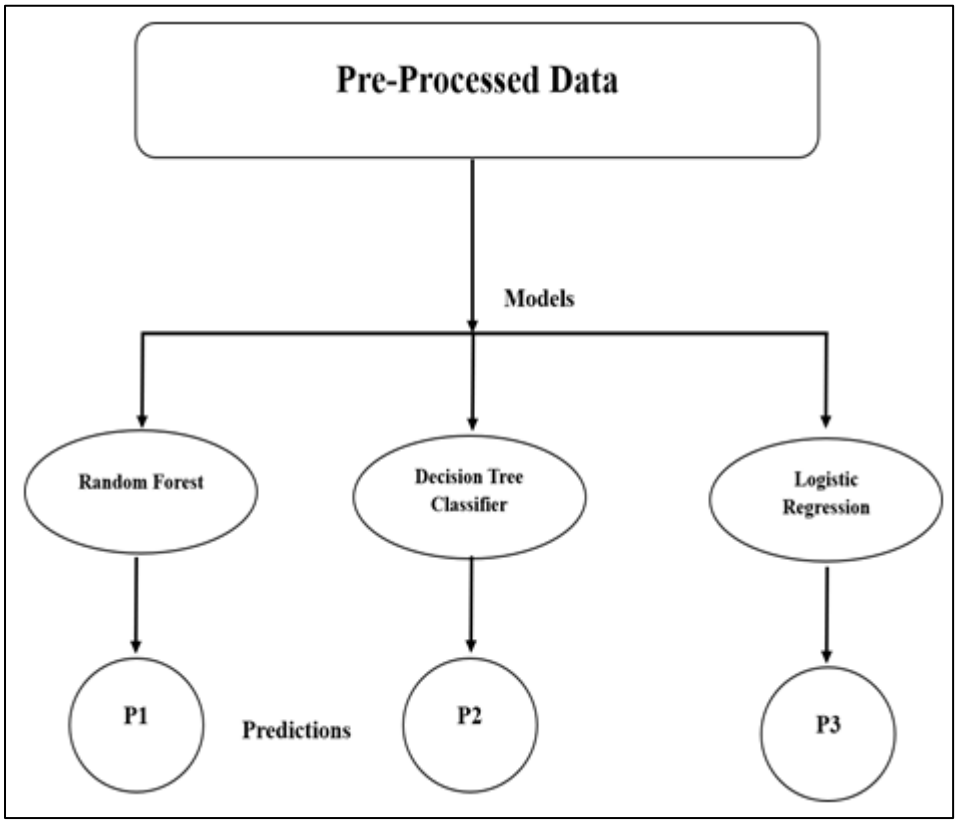

Fig 1: Basic Workflow of Model

A. Workflow Architecture

- User Interaction: The user accesses the web interface (index.html) and inputs their data. The user selects a machine learning model for prediction.
- Data Submission: The form data is submitted to the Flask backend via a POST request.
- Data Preprocessing: The backend preprocesses the input data by encoding categorical features and scaling numeric features.
- Model Prediction: The backend loads the selected model from the .pkl file and makes a prediction. The prediction result is returned to the user via the web inter- face.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14737958

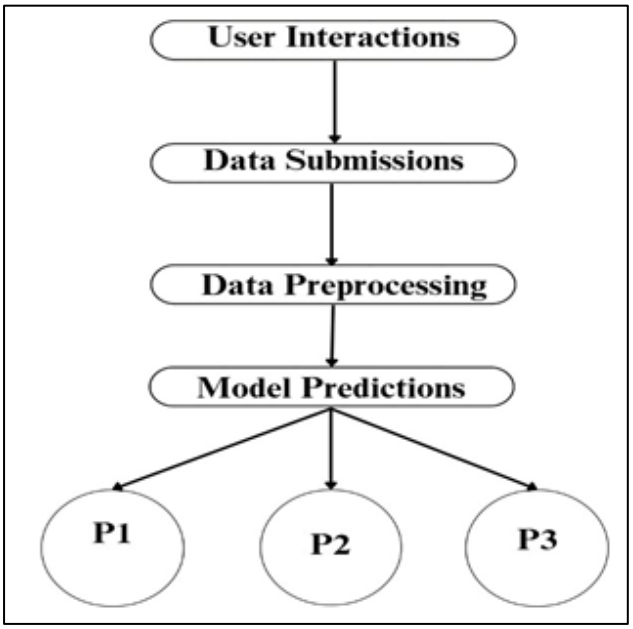

Fig 2: Decision Tree Results

B. Data Flow

- User Input: The user inputs their data via the web interface.
- Data Submission: The form data is submitted to the Flask backend.
- Data Preprocessing: The backend preprocesses the input data.
- Model Prediction: The backend loads the selected model and makes a prediction.
- Result Display: The prediction result is displayed to the user via the web inter- face.

VI. DATASET DESCRIPTION

The dataset has 27,902 entries and 11 variables, including demographic, lifestyle, academic, and mental health information. Important variables are Gender (male/female), Age (18-42), Academic Pressure (1-5 scale), Satisfaction towards studies (1-5 scale), Du- ration of sleep (minutes, such as ”less than 5 hours”), Eating habits (Healthy / Moderate/ Unhealthy), Suicidal thoughts (Yes/No), Financial pressure (1-5 scale), Family history of mental illness (Yes/No), and Depression (0/1). It is quite reasonable to investigate the interplay between lifestyle factors (for instance, sleep, diet) in relation to mental health issues (for instance, depression, suicidal thoughts). Key analyses include correlation studies (for example, correlation between sleep duration and depression), prediction modeling (for example, predicting depression in terms of academic pressure and financial stress), and demographic insights (for example, age/gender differences in mental health). Questions such as, ”Does higher academic pressure correlate with suicidal thoughts?” or ”Can healthy habits have effects on study satisfaction?” would be addressed. The dataset is useful for categorical, regression, and statistical analyses; thus, it becomes quite useful for mental health research and identification of risk factors. However, feel free to ask for specific analyses or visualizations, if desired.

VII. RESULTS AND EVALUATION

Logistic Regression trained.

	precision	recall	f1-score	support
0	0.83	0.77	0.80	3482
1	0.85	0.89	0.87	4888
accuracy			0.84	8370
macro avg	0.84	0.83	0.83	8370
weighted avg	0.84	0.84	0.84	8370

Fig 3: Logistic Regression Results

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14737958

Decision Tree trained.

	precision	recall	f1-score	support
0	0.81	0.80	0.80	3482
1	0.86	0.86	0.86	4888
accuracy			0.84	8370
macro avg	0.83	0.83	0.83	8370
weighted avg	0.84	0.84	0.84	8370

Fig 4: Decision Tree Results

Random Forest trained.

	precision	recall	f1-score	support
0	0.82	0.78	0.80	3482
1	0.85	0.88	0.86	4888
accuracy			0.84	8370
macro avg	0.83	0.83	0.83	8370
weighted avg	0.84	0.84	0.84	8370

Fig 5: Random Forest Results

VIII. CONCLUSION

This approach can assist educators, counselors, and legislators in taking prompt action that benefits kids' academic and personal lives by using a machine learning model to accurately anticipate students' sadness.

Using an ensemble technique has improved prediction reliability by combining the advantages of several machine learning models. Designing focused actions to improve a healthy academic environment will be made possible by the system's insights into the main impacting factors.[7]

By adding more dimensions and including real-time data streams for ongoing monitoring and forecasting, future research may be able to access a larger dataset from various locations. This research illustrates the multidisciplinary applicability of machine learning as well as its capacity to lead social change.[8]

REFERENCES

[1]. et al. Paulo Mann. Detecting depression symptoms in higher education students using multimodal social media data. *arxiv.org*, 2020.

[2]. et al. Radwan Qasrawi. Assessment and prediction of depression and anxiety risk factors in schoolchildren: Machine learning techniques performance analysis. *JMIR FORMATIVE RESEARCH*, 2022.

[3]. et al. Ayako Baba1. Prediction of mental health problem using annual student healthsurvey: Machine learning approach. *JMIR MENTAL HEALTH*, 2023.

[4]. Narasappa Kumaraswamy. Academic stress, anxiety and depression among college students- a brief review. *International Review of Social Sciences and Humanities*, 2012.

[5]. et al. Nguyen M.-H. A dataset of students' mental health and help-seeking behaviors in a multicultural environment. *MDPI*, 2019.

[6]. et al. Cai H. A pervasive approach to eeg-based depression detection. *Wiley Com- plexity*, 2018.

[7]. et al. Jiang T. Addressing measurement error in random forests using quantitative bias analysis. *American Journal of Epidemiology*, 2021.

[8]. et al. Lebedev A. Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness. 2014.

[9]. et al. Cacheda F. Early detection of depression: social network analysis and random forest techniques. *Medical Internet research*, 2019.

[10]. Sau A. and Bhakta I. Artificial neural network (ann) model to predict depression among geriatric population at a slum in kolkata, india. *Journal of clinical and diagnostic research: JCDR*, 2017.

[11]. et al. Wade B.S. Random forest classification of depression status based on subcor- tical brain morphometry following electroconvulsive therapy. 2015.

[12]. Garg S. Priya A. and Tigga N.P. Predicting anxiety, depression and stress in modern life using machine learning algorithms. 2020.

[13]. et al. Islam M.R. Depression detection from social network data using machine learning techniques. 2018.

[14]. et al. Supriya S. Eeg sleep stages analysis and classification based on weighed complex network features. 2018.

[15]. Mohanavalli S. Srividya M. and Bhalaji N. Behavioral modeling for mental health using machine learning algorithms. *Journal of medical systems*, 2018.

[16]. et al. Pflueger M.O. Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. 2015.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14737958

[17]. et al. Banda J.M. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. 2019.
[18]. et al. Laijawala, V. Mental health prediction using data mining: A systematic review. 2020.
[19]. et al. Chutia D. An effective ensemble classification framework using random forests and a correlation based feature selection technique. 2017.
[20]. Nithya B. and Ilango V. Predictive analytics in health care using machine learning tools and techniques. 2017.