Detection of depression in Students Using Machine Learning Models

Dr. Shobha T[1], Sreevidya B S[2], Manjula S[3]

[1]Assistant Professor, Department of CSE, BMSCE

[2]Assistant Professor, Department of CSE, BMSCE

[3]Assistant Professor, Department of ISE, BMSCE

Abstract

Many millions of individuals experience mental diseases that are frequently ignored, and a large portion of these people are children and young adults, many of whom are students. Mental health has increasingly changed into a field that has sparked interest in almost every field and gained attention. This study identified 5 techniques of machine learning and evaluated their efficacy in detecting student depression using a number of accuracy metrics. The 5 methods of machine learning used in this study are random forest, logistic regression, gradient boosting, SVM and decision tree classifier. The comparison of these methods put them into practice and found that Random Forest method to be the most accurate among other classifiers, with a prediction accuracy of 85%.

Keywords- Machine Learning, Depression Detection, Mental Health, Students, SVM, decision tree classifier, random forest, logistic regression, gradient boosting

I. INTRODUCTION

The effects of depression on people's psychology are severe. Depression will thus impair a person's capacity to focus, study, and perform efficiently, which will have a significant negative influence on the individuals' daily life. Depression and suicidal thoughts are closely related, and depression itself can result in suicide. People with depression frequently face stigma from the public and social exclusion from their families. They could not perform as well in workplaces and educational institutions. People are consequently losing access to economic and social opportunities, which has a detrimental effect on their quality of life. Five of the highest 10 most common diseases in the world that render people disabled or incapable are mental illnesses, with depression coming in at number one. This is sufficient evidence that depression can seriously hurt society. Currently, a questionnaire survey is the primary method for diagnosing depression. In general, psychiatrists utilize questionnaires to evaluate depression because individuals with depression are frequently reluctant to discuss their emotions with professionals, family members, or friends and prefer to do it anonymously. Therefore, without consulting any medical professionals or engaging in face-to-face interaction, an effort to evaluate the depth of a depression using computers. The data was obtained from www.kaggle.com and was gathered through online questionnaires that various students completed.

The large amount of information that will be available for medical professionals and mental health researchers is by smart technologies like neuroimaging, Smartphones, social media, and wearable technology have made it possible for medical professionals. The growing technology of Machine Learning

(ML) facilitated them to analyze these data. Advanced Probabilistic and Statistical (P & S) methods utilized in ML to build models that can independently learn from data. This makes it possible to predict results from data sources as much as accurately and correctly identify data trends. The analysis of mental health data using ML methods has the potential to enhance patient outcomes as well as knowledge about psychiatric illnesses and how to treat them.

II. RELATED WORK

Research related to mental health diagnosis using machine learning techniques began way back in the 1980's.To diagnose depression, many data collection methods and machine learning algorithms have been explored. Data acquired by questionnaires are more precise and specific. Ms. Sumathi M.R. and Dr. B. Poorna [1] manually generated a data set by identifying twenty-five attributes along with class labels (level of depression), the number of data instances was just sixty. This model can be improved by using more data examples;[2] the dataset with 27 columns and 1259 items was used to train 5 different machine learning methods, all of which resulted in accuracy greater than 79%. Based on existing questionnaires such as the DASS-21 (Depression Anxiety And Stress Scale), GAD-7 (General Anxiety Disorder), and with the assistance of a psychologist, the authors[6] designed novel questionnaires comprising of 40 questions and distributed them to university engineering students. A total of 127 replies were received. Random forest algorithm performed best, yielding an accuracy of 78.9%.

Life situations that are stressful typically have an adverse effect on a person's mental well-being. Loneliness and unhappiness might contribute to the growth of depression. Bo Lin ,Junbo Ma [3] conducted a survey involving 1659 college students to explore the relationship between stressful life events, loneliness, and depression. Loneliness was identified as a mediator, meaning it indirectly linked stressful events to depression. Various other features like demographic information, aspects related to university life, physical health problems, and relationships with family and friends were included in the dataset by the authors [4].The main drawback was the imbalance distribution between non depression and depression samples in their study.

The emergence of social media platforms such as Facebook, Instagram and Twitter not only offers more easy contact options for students and also gives a new emotional venting window for students. The students can use social media like named above to document their living situations in real time and engage with others to express feelings (happy, sad, emotional etc.,) and reduce stress. Simultaneously, the evolution of social media has created a new method for detecting sad people. To diagnose depression, contemporary computer technology analyzes the user's data from social networks. In paper [9], the study employs text mining of Sina Weibo data. Following the extraction of properties by deep neural networks, the Deep Integrated SVM (DISVM) technique classifies the data for depression detection. The suggested method efficiently locates possible depression sufferers utilizing Sina Weibo data. The authors [10] employ two separate datasets, one from a questionnaire similar to the PHQ-9 (Patient Health Questionnaire) and the other from twitter and reddit posts using NLTK. For dataset I, the XGBoost classifier had the greatest accuracy of 83.87%, while Logistic regression had the maximum accuracy of 86.45% for dataset II.

Depression detection using ML is not just limited to questionnaire dataset, authors [5] proposed an intelligent social therapeutic chatbot wherein it labels the emotion namely, Happy, Joy, Shame, Anger, Disgust, Sadness, Guilt, and Fear and distributes the text into these emotion labels. Further, based on these emotion label, it recognizes the users' mental state such as depression, stress using users' chat or posted data. To detect emotions, they employed three Deep Learning classifiers : Recurrent Neural

Network (RNN), Convolutional Neural Network (CNN) and Hierarchical Attention Network (HAN).

Social learning provides a multidimensional data set that contains a massive information that can express user's behavior, alike speech subject information and emotional information. These traits are also efficient for detecting users' psychological states. Based on voice fragment traits, this study [7] employed machine learning to identify depression in Chinese students. Bi-LSTM and CNN networks were used to extract features from speech fragments and process them. An attention mechanism was used to choose depression-related features, and a Full Connection layer was used to forecast depressive tendencies.

Facial expression detection is used by researchers and scientists to study the topic of emotion recognition. In Paper [8], authors presented the Depression Possibility Detection Tool (DPDT) as an automated tool for assessing depression probability in students. DPDT employs facial expressions, eye movements, behavior changes, and physical conditions as indicators

III. IMPLEMENTATION

Before the system is able to accurately predict the end result, it goes through a number of steps. Data collection, data preprocessing and feature selection, choosing training and test data, and applying ML algorithms are these processes as shown in Figure 1. After achieving the requisite accuracy, and can combine the system with an application for use in the real world

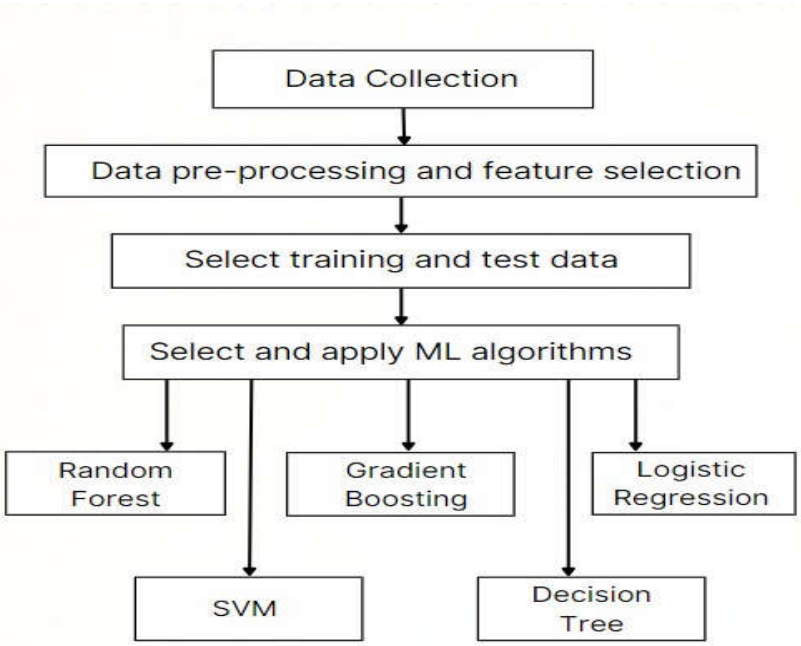

Figure 1: Flow Diagram of the methodology

a) Data Collection

The dataset was obtained from Kaggle. It is made up of 18 questions and 352 instances that were gathered from surveys of students enrolled in various levels of education, namely high school, college, bachelor's, and master's programmes.

b) Data pre-processing and Feature selection

There were initially 18 features to foresee the outcome. The predicted performance of a learning

system may decline due to collinearity and interaction within the features. Therefore, lets look for the covariance matrix to identify the redundant features. It provides details on the correlation of feature movement. Finally, the best suited collection of features was found and utilized when performing predictive modeling.

The study also used feature importance along with correlation matrix for feature selection. This demonstrates the contribution of each feature to the model's forecast. Table 1 shows the feature scores of different features

Aspect	Value
Education	3251.247506
Interest	8446.462181
Sleep	6990.412265
Energy	10875.48503
Appetite	8089.683368
Self_Worth	20383.05203
Concentration	8457.327726
Restlessness	6215.412411
Suicidal_Thoughts	9390.922844
Job	4651.829326
Accommodation	3930.938179
Study_Hours	4278.448489
Social_Media_Hours	5038.778647

Table 1. Feature importance score

c. *Classification*

Using Colab notebooks and the Python programming language, machine learning methods were applied. This predicts the proportion of people who experience depression. The training and test sets of the dataset were partitioned in the ratio 80:20, respectively. The information along with labels is used for the training and testing of the classifiers. Random forest, decision trees, logistic regression, SVM and gradient boosting classifiers are applied on the data.

Finding the algorithm with the highest level of accuracy for diagnosing depression is the last step. The percentage of accurately predicted classes divided by the total number of predictions can be used to determine accuracy. Each performance measure for categorization is subject to the calculation. Once the flow has stopped, the most accurate method for detecting depression will be determined.

Confusion Matrix is a method for analyzing how well a particular algorithm performed in order to determine the best classifier for a given issue set. Here. proposed a concept to identify the method with the best accuracy thanks to the confusion matrix. The actual class is displayed in the confusion matrix's rows, while the anticipated class is displayed in its columns. Figure 2. shows the prediction summary of the random forest classifier

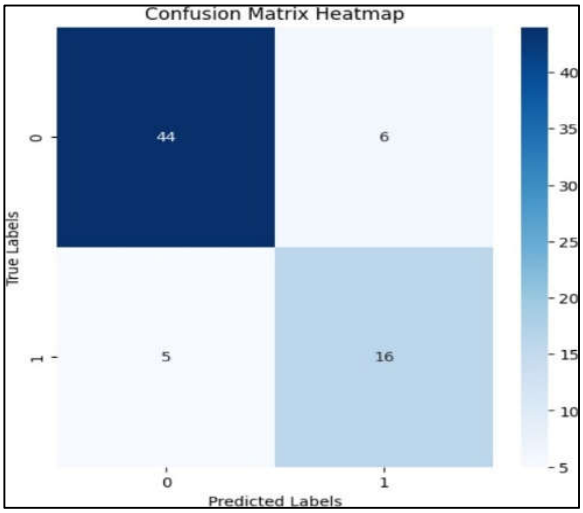

Figure 2. Confusion matrix

IV. RESULTS

The data must be adequately cleaned and preprocessed until excellent accuracy can be attained by efficiently fitting the model. This study identified five machine learning methods: random forest, decision trees, logisticregression, SVM and gradient boosting. Evaluated how well they did at spotting student depression. Table 2 displays the various classification algorithms' accuracy, precision, recall, and f1 scores. According to the findings of Table 2, Random Forest achieved the best accuracy of 85%.

Algorithm	Accuracy	Precision	Recall	F1 score
Decision Tree	0.81	0.88	0.85	0.87
Logistic Regression	0.79	0.90	0.78	0.84
SVM	0.81	0.88	0.83	0.86
Random Forest	0.85	0.91	0.86	0.89
Gradient Boosting	0.81	0.88	0.83	0.86

Table 2: Values of multiple metrics for various classification techniques

V. CONCLUSION.

In a time of escalating competitiveness, more students are experiencing depression. In this paper, a technique for quickly identifying depressed students is suggested. It is crucial to examine the various machine learning approaches currently accessible to be able to choose the one that matches the target domain. To ensure the treatment is carried out successfully and efficiently, a number of specialized programmes are now available in the healthcare sector that can forecast diseases quite precisely in future. In the suggested work, the dataset was categorized utilizing 5 distinct machine learning approaches. The algorithm that produced the best results, with an accuracy of 85%, was random forest. This study may be employed with bigger datasets in the near future for increased accuracy. The research only used a relatively small dataset. Investigate cutting-edge ML methods, such as deep learning

architectures, to improve the model's capacity to detect subtle patterns and connections in the dataset..

REFERENCES

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

[2] Vaishnavi, K., Kamath, U. N., Rao, B. A., & Reddy, N. S. (2022). Predicting mental health illness using machine learning algorithms. In Journal of Physics: Conference Series (Vol. 2161, No. 1, p. 012021). IOP Publishing.

[3] Lin, B., & Ma, J. (2021, July). The Effect of Stressful Life Events on College Students' Depression: A Moderated Mediating Model. In 2021 International Conference on Public Health and Data Science (ICPHDS) (pp. 178-183). IEEE.

[4] Nison, P., Vuttipittayamongkol, P., Boonyapuk, P., & Kemavuthanon, K. (2023, January). A Machine Learning Approach for Depression Screening in CollegeStudents Based on Non-Clinical Information. In 2023 International Conference On Cyber Management And Engineering (CyMaEn) (pp. 413-417). IEEE.

[5] Patel, F., Thakore, R., Nandwani, I., & Bharti, S. K.(2019, December). Combating depression in students using an intelligent chatBot: a cognitive behavioral therapy. In 2019 IEEE 16th India Council International Conference (INDICON) (pp. 1-4). IEEE.

[6] Bhatnagar, S., Agarwal, J., & Sharma, O. R. (2023).Detection and classification of anxiety in university students through the application of machine learning. Procedia Computer Science, 218, 1542-1550.

[7] Qu, M., Lu, X., Liu, Y., Pan, T., Liu, J., & Wang, Y.(2023, March). Depression recognition in university students based on speech features in social learning environment. In 2023 International Conference on Artificial Intelligence and Education (ICAIE) (pp. 30-34). IEEE.

[8] Gamage, M. A., Arachchi, R. M., Naotunna, S., Rubasinghe, T., Silva, C., & Siriwardana, S. (2021,December). Academic Depression Detection Using Behavioral Aspects for Sri Lankan University Students.In 2021 3rd International Conference on Advancementsin Computing (ICAC) (pp. 335-340). IEEE.

[9] Ding, Y., Chen, X., Fu, Q., & Zhong, S. (2020). A depression recognition method for college students using deep integrated support vector algorithm. IEEE access, 8, 75616-75629.

[10] Jain, S., Narayan, S. P., Dewang, R. K., Bhartiya, U., Meena, N., & Kumar, V. (2019, May). A machine learning based depression analysis and suicidal ideation detection system using questionnaires and twitter. In 2019 IEEE students conference on engineering and systems (SCES) (pp. 1-6). IEEE.

[11] Govindasamy, Kuhaneswaran AL, and Naveen Palanichamy. "Depression detection using machine learning techniques on twitter data." In 2021 5th international conference on intelligent computing and control systems (ICICCS), pp. 960-966. IEEE, 2021.

[12] Al Asad, Nafiz, Md Appel Mahmud Pranto, Sadia Afreen, and Md Maynul Islam. "Depression detection by analyzing social media posts of user." In 2019 IEEE international conference on signal processing, information, communication & systems (SPICSCON), pp. 13-17. IEEE, 2019.

[13] Hassan, Anees Ul, Jamil Hussain, Musarrat Hussain, Muhammad Sadiq, and Sungyoung Lee. "Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression." In 2017 international conference on information and communication technology convergence (ICTC), pp. 138-140. IEEE,2017.