Depression Detection Using Machine Learning

[1] Sahana Hegde
PG Scholar, Department of Computer Applications
Dayananda Sagar College of Engineering
Bangalore, India
sahanahegde22@gmail.com

[2] Mahendra Kumar B
Associate Professor, Department of Computer Applications
Dayananda Sagar College of Engineering
Bangalore, India
mahendra-mcavtu@dayanandasagar.edu

***Abstract*— A mental disease called depression can negatively affect your thoughts and behavior. These days, depression is a frequent condition. A potential area of research involves using machine learning to identify depression using data from social media, medical records, and physiological signs. Using supervised approaches like Supervised machine learning, decision trees, and neural networks, people are categorized as having depression or not. The population of people with depression is divided into subgroups using unsupervised techniques, such as clustering. Methods for feature extraction and selection point out pertinent features. Utilizing performance criteria like accuracy, precision, recall, and AUC-ROC, model efficacy is evaluated. By incorporating machine learning into healthcare, depression can be detected early and treated individually. Data quality, privacy, and interpretability are issues. Clinical trials and prospective studies are essential to confirming real-world efficacy.**

***Keywords*— Depression Detection;Machine Learning; Supervised Algorithm;Random Forest;Naive Bayes;Feature Selection.**

I. INTRODUCTION

Depression is a common mental health disorder that affects millions of individuals worldwide and is extremely expensive, socially isolating, and emotionally draining. Effective treatment and support depend on early recognition and intervention. Traditional approaches of diagnosing depression place a significant emphasis on subjective evaluations, which can be impacted by a number of circumstances. As a result, machine learning techniques are being used increasingly frequently to improve the accuracy and efficacy of diagnosing depression. The "machine learning" sector of artificial intelligence is concerned with developing algorithms as well as models which can recognize patterns in data, forecast or categorize events, and learn from it. The present research analyses the utilization of data specifically collected from students to identify sadness using machine learning methods. The goal of the project is to "discover pattern and develop predictive models that can identify individuals at risk of depression." It does this by examining several factors associated to anxiety and depression among students. The use of machine learning to diagnose depression has already yielded promising results. Numerous datasets have been used in studies to discover important characteristics and patterns suggestive of depression, including clinical data, social media posts, and survey responses. However, there is a need to look into how to identify depression among students specifically, as they have certain stressors and difficulties that may contribute to the onset of depressed symptoms.

The dataset used in this study, the "Students Anxiety and Depression Dataset," provides a comprehensive collection of information related to anxiety and depression symptoms among students. It includes demographic data, academic performance, social interactions, lifestyle factors, and self-reported anxiety and depression scores. By leveraging this dataset, the study aims to develop machine learning models that can accurately detect depression among students. The methodology employed in this research paper involves several essential steps. Initially, the dataset undergoes a preprocessing stage to handle missing values, normalize features, and address any data imbalances. Then, using feature selection approaches, the most pertinent traits that considerably aid in depression diagnosis are found. The effectiveness of different approaches to machine learning, including random forests and naive bays, is evaluated through training and evaluation using the right criteria. The predicted outcomes of this work include the creation of machine learning models that, using the available dataset, can accurately identify depression in students. We will measure the results of the models and compare them with existing approaches using common assessment metrics, including accuracy and precision, recall scores. Additionally, by concentrating exclusively on students—a vulnerable population with particular needs and characteristics—the study hopes to add to the body of knowledge on depression detection. The examination into depression identification using machine learning approaches is presented in this research paper's conclusion. The goal of the study is to create precise and effective models for diagnosing depression in students using the "Students Anxiety and Depression Dataset." The outcomes and understandings from this research can help early intervention and support systems, ultimately enhancing student wellbeing and tackling the global depression epidemic.

Furthermore this paper has covered literature survey on some previous papers in the area and post to that,by considering a dataset different algorithms have been implemented and furthermore the results have been compared ,post to that we come to a conclusion based on the observing and results.

II. LITERATURE SURVEY

This part will cover a review of depression, current methods for detecting depression using different machine learning algorithms, and how to identify holes in the body of existing research and fill them in the proposed study. The section is broken up into a number of sections.

Researchers have made the case in a review article titled "Accuracy of Machine Learning in the Depression Detection" that depression is a major mental health issue that, if disregarded, can have serious repercussions. However, detecting depression is challenging due to factors such as patient denial and societal stigma. Social media posts can exhibit symptoms of depression, making them a potential source for detection using machine learning algorithms. These algorithms serve as an alternative or support for psychologists in diagnosing depression, providing accurate results depending on the dataset. To identify the most accurate machine learning algorithm for depression detection, a systematic literature review was conducted. The findings indicate that Logistic Regression achieves the highest accuracy of 99.80%. The accuracy of dependable algorithms like Logistic Regression and Support Vector Machines (SVM) is constantly over 70%. The Twitter dataset is the one that has been used most frequently in earlier studies. In general, machine learning techniques show promise in precisely identifying depression, with Logistic Regression appearing as a very accurate choice. [2]

A different scholar by the name of Nafiz Al Asad has tried with detecting depression through investigating user-posted social media content.[10]

A person's manner of life might be affected by significant depressive illnesses. Negative impact and mood are caused by depression. Individuals began to miss those past pastimes as a result of their impact on them.[3][4]

In a 2020 research study titled "Detection of the Depression in the Social Media via Twitter using Machine Learning Approach," the authors have utilized the naive bayes method to draw a conclusion
from data collected via social networks.[6]

- Doctors are able to do procedures remotely because to monitors that can show information about the heartbeat, pulse, and other vital signs. Even if there hasn't been great strides regarding the management of mental disorders, scientists and psychologists are collaborating with different tech businesses to find the kinds of digital tools, software that are certain to assist with this sickness. The state of sad persons worldwide is explained by a number of statistics. [7]
- More than 264 million individuals globally are afflicted by mental disease. combining sufferers of depression ranging from profound to moderate.[7]
- Every year, over 8,000,000 people kill themselves. With suicide ranking as The next most frequent cause of death among individuals between the ages of 15 and 29.[7]
- Between 76% and 85% of persons in nations with low to middle incomes are not getting any sort of therapy for mental problems.[7]
- After evaluating data from the National Survey of Children's Health from the year of 2016, it was shown that 3.2% of kids aged 3 to 17 in the US had depression, 7.4% had behavioral conduct issues, and 7.1% had anxiety issues. 4.4 ,4.5 million youngsters, respectively, could be included in these.[8]
- After carefully examining the aforementioned data, it was shown that children with depression received roughly 20% fewer medications than those with anxiety.[8]
- Approximately 78% of kids who have depression have had therapy, 59% of kids who have anxiety, and 54% of kids with behavioral conduct issues.[8]

Students must develop coping mechanisms and abilities in a classroom environment in order to handle mental illness among them. The hiring of counsellors and psychologists should also be given priority in schools, rather than spending a lot of money on ostentatious infrastructure and self-promotion.

III. RESEARCH METHODOLOGY

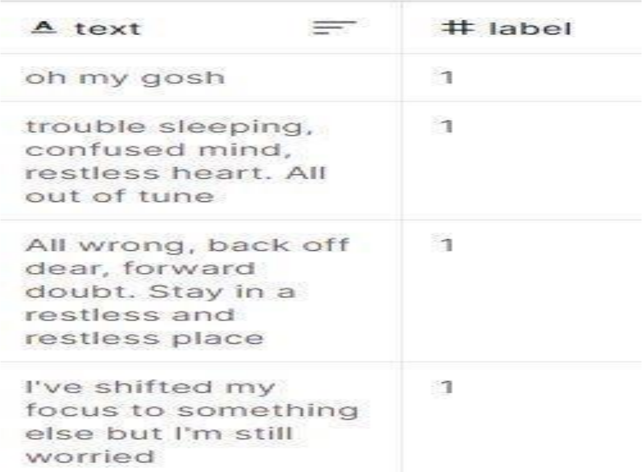
The gathering of data is the first step in the process. The "students-anxiety-and-depression-dataset" dataset already exists and contains information gathered from numerous anonymous students and saved in an .xslx file. Data pre-processing is the process' next phase. Here, pre-processing includes several processes like stemming and tokenizing.

Two distinct algorithms are used to process the data: 1) Naive Bayes 2) Random Forest

The train set and test set of data are separated. To ensure the classifier learns, the model is developed using the training data. When the model has gained knowledge about the data for evaluation, the test data will be fed into the model.[1]

A. Data Collection

Kaggle was used to acquire around 7000 data points in total.



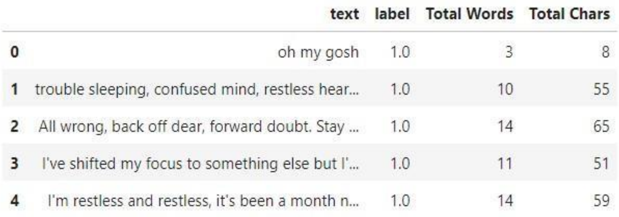
A text	# label
oh my gosh	1
trouble sleeping, confused mind, restless heart. All out of tune	1
All wrong, back off dear, forward doubt. Stay in a restless and restless place	1
I've shifted my focus to something else but I'm still worried	1

Fig. 1. Example of Raw Dataset

People can search for datasets on Kaggle that they are interested in using to develop artificial intelligence (AI) simulations, publish datasets, partner with other data scientists and machine learning specialists, and take part in initiatives that tackle data science issues. When ready for evaluation, the data is subsequently saved in an .xslx file. The dataset includes the columns
"text" and "label," respectively.

B. Data pre-processing

The dataset was run on Kaggle it self for pre-processing.



	text	label	Total Words	Total Chars
0	oh my gosh	1.0	3	8
1	trouble sleeping, confused mind, restless hear...	1.0	10	55
2	All wrong, back off dear, forward doubt. Stay ...	1.0	14	65
3	I've shifted my focus to something else but I'...	1.0	11	51
4	I'm restless and restless, it's been a month n...	1.0	14	59

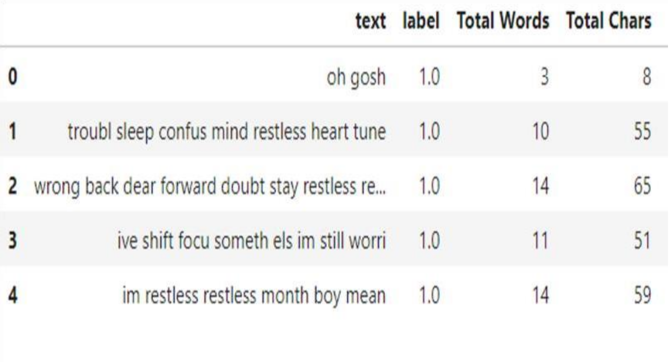
Fig. 2. Example of Unprocessed Data

Punctuation, stopwords, URLs, and other elements were eliminated from the dataset. The sentence is then divided into each token or word in an array format, and the dataset is tokenized as a result. Sentences that have been tokenized are next subjected to lemmatization and stemming. Stemming is a technique for removing affixes from words so that the original word can be recovered. [1]

For instance, the word "Running" will become "Run" in normalization.

Lemmatization, in essence, comprehends the phrases and returns the term that signifies in some circumstances stemming could result in inaccurate meanings and spelling mistakes. Lemmatization, however, corrects its and reverts to the useful basic form. While processing the data with the help of split and length functions, the number of characters and words are counted. And total words and total characters are added as new columns.

This is what processed data looks like-



	text	label	Total Words	Total Chars
0	oh gosh	1.0	3	8
1	troubl sleep confus mind restless heart tune	1.0	10	55
2	wrong back dear forward doubt stay restless re...	1.0	14	65
3	ive shift focu someth els im still worri	1.0	11	51
4	im restless restless month boy mean	1.0	14	59

Fig. 3. Example of processed Data

A. Classification-

a) Naive Bayes: Naive Bayes, also referred to as NB, is a classification-related supervised machine learning algorithm.

Describe categorization. - The Classification method is a Supervised Learning technique that classifies new findings based on data from training. The model is asked to determine the proper label for given input data using the supervised machine learning technique known as classification.

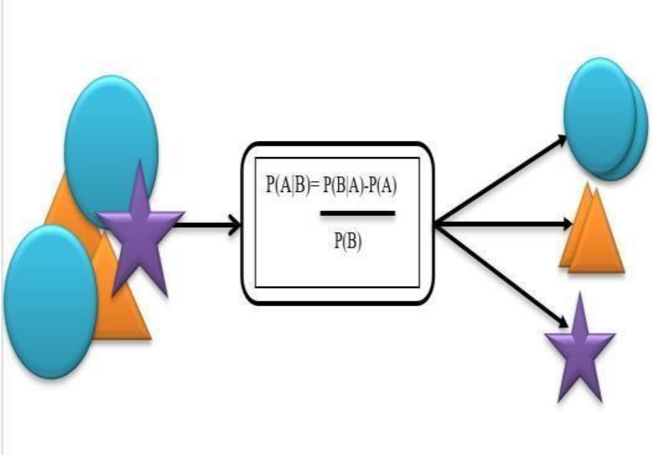


Fig. 4. Example of Naïve Byes Classifier

Here, NB employs the same process. Mostly, it has to do with categorizing texts. Due to the fact that it is based on forecasts of an object's probability, it is referred to as a probabilistic classifier.

Why is it called "Naive Bayes"?:

It is recognized as naive because it assumes that the presence of one feature is unrelated to the presence of another feature. Because it is based on the Baye's theorem principle, it is known as a Bayes model.

$$P(B|A) = \frac{P(A|B)*P(A)}{P(B)}$$

b) Random Forest: A popular supervised machine learning method is Random Forest. In an effort of boosting the predicted quality of the data set being used, the Random Forest model combines the results from several decision trees that were applied to multiple subgroups of the input sample. The forest is composed of varieties of tree species, and also the greater the distinct tree species, the better the forest is going to be. The algorithm's correctness and problem-solving capacity rise in a manner identical to it as the amount of trees grows. For the purpose of boosting the projected quality of the dataset, a classifier called as Random Forest uses several decision trees on different portion of the information that is provided. It relies on the notion of ensemble learning, which is the activity of mixing multiple classifiers to deal with challenging issues and enhance the effectiveness of the framework.

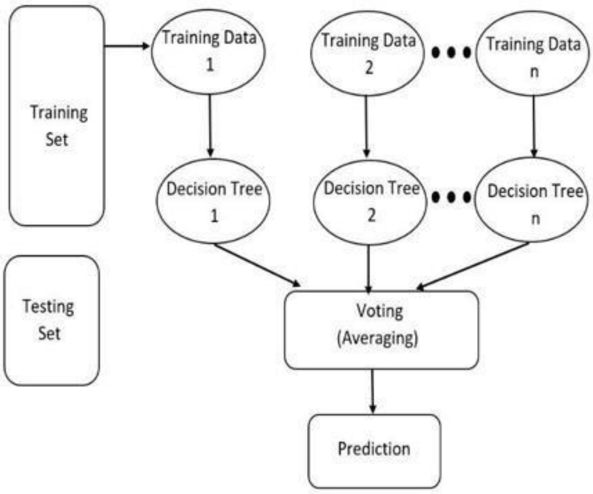


Fig. 5. Random Forest Working

Decision trees are the only source of data used by Random Forest. The name "Random Forest" refers to the fact that a forest with more trees will be more accurate. The outcome is predicted by the random The following steps of the Random Forest.

Algorithm's functioning are outlined:

Step 1- select random samples from an existing data collection or training set.

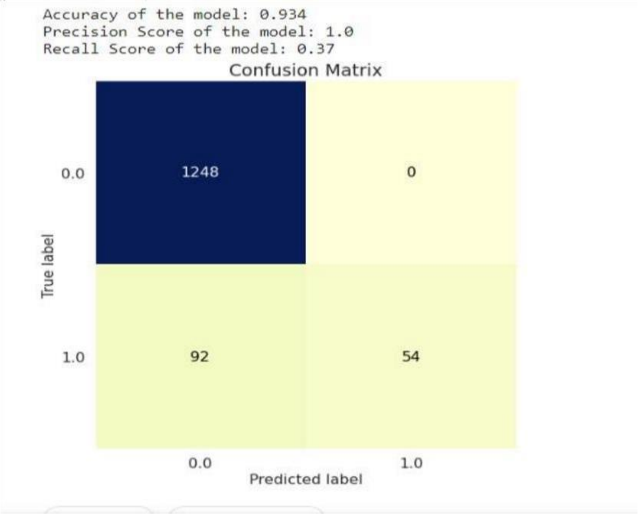
Step 2-The second step involves this algorithm creating a decision tree for each training set of data.

Step 3-The decision tree's average will be utilized for the polling.

Step 4- The outcome with the most consensus should be chosen as the ultimate expected outcome.

IV. RESULTS AND COMPARISION

Here, we examine and contrast the outcomes of the two methods we've used to produce them. Forest using wisdom from each tree and the majority of choices made.



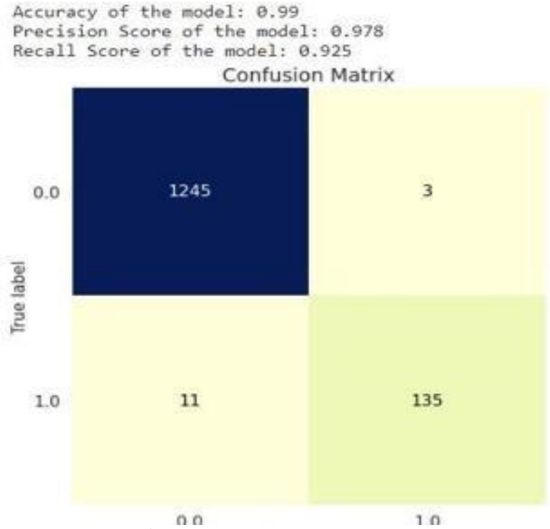
```
Accuracy of the model: 0.934
Precision Score of the model: 1.0
Recall Score of the model: 0.37
```

Confusion Matrix

True label \ Predicted label	0.0	1.0
0.0	1248	0
1.0	92	54

Fig. 6. Result of Naïve Bayes

The Naïve Bayes achieves 93.4% of success. That means 93.4% of the data has been properly categorized.



```
Accuracy of the model: 0.99
Precision Score of the model: 0.978
Recall Score of the model: 0.925
```

Confusion Matrix

True label	0.0	1.0
0.0	1245	3
1.0	11	135

Fig. 7. Findings from Random Forest

Similarly Random Forest Algorithm achieves 99% of success. That means 99% of data has been considered dataset Random Forest shows 5.6% high accuracy.

Comparison :

Algorithm	***Naïve Byes***	***Random Forest***
Accuracy	93.4	99

TABLE I. COMPARISON TABLE

The above table shows the results obtained . For the considered dataset Random Forest algorithm shows 5.6% high accuracy.

V. CONCLUSION

In conclusion, this research project looked at ways to recognize sorrow in students using machine learning techniques. Depression is a prevalent mental health illness that negatively affects a person's social, economic, and personal lives. Early detection and intervention are essential for effective treatment and support. Conventional methods for diagnosing depression, however, mostly rely on subjective assessments that may be influenced by a variety of variables. Thus, machine learning is increasingly being used to increase the accuracy and efficacy of depression identification. The goal in here was to study accurate and efficient machine learning models for identifying depression in students using the "Students Anxiety and Depression Dataset." The extensive information in the dataset included demographic information, academic performance, social interactions, lifestyle traits, and self-reported anxiety and depression scores. Through a number of stages, such as data preprocessing, feature selection, model training, and evaluation, the study set out to identify trends and develop prediction models that could identify individuals who are at risk of developing depression. One of the expected objectives of this work is the development of machine learning models that can accurately detect student depression using the existing dataset. To enable comparison with current techniques, these models will be assessed using accepted performance standards like recall and accuracy. Furthermore, the study aimed to contribute to the field of depression diagnosis by focusing especially on students, a susceptible population with unique stressors and challenges. The information gathered from this study might be used to create early intervention and support initiatives that would improve student welfare and address the global depression epidemic. To confirm the viability of the suggested machine learning models in the actual world, additional study and clinical trials are required.

In summary, this study looked into applying machine learning to detect student depression. By leveraging a huge dataset and a range of methodologies, the study aimed to develop precise models that can enhance early identification and individualized treatment of depression, addressing an important area of mental health care.

REFERENCES

[1] K. A. Govindasamy and N. Palanichamy, "Depression Detection Using Machine Learning Techniques on Twitter Data," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 960966, doi: 10.1109/ICICCS51141.2021.9432203.

[2] A. M. Putri, K. Wijaya, O. A. Salomo, A. A. Santoso Gunawan and Anderies, "A Review Paper: Accuracy of Machine Learning for Depression Detection in Social Media," 2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), Solo, Indonesia, 2022, pp. 39-45, doi: 10.1109/COMNETSAT56033.2022.9994553.

[3] M. Hooda, A. R. Saxena, D. Madhulika and B. Yadav, "A Study and Comparison of Prediction Algorithms for Depression Detection among Millennials: A Machine Learning Approach," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, India, 2017, pp. 779-783, doi: 10.1109/CTCEEC.2017.8455078.

[4] Blanco, Joel A., and Lynn A. Barnett. "The effects of depression on leisure: varying relationships between enjoyment, sociability, participation, and desired outcomes in college students." Leisure Sciences 36.5 (2014): 458-478

[5] H. Sanyal, S. Shukla and R. Agrawal, "Study of Depression Detection using Deep Learning," 2021 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2021, pp. 1-5, doi: 10.1109/ICCE50685.2021.9427624.

[6] P. Kumar, R. Chauhan, T. Stephan, A. Shankar and S. Thakur, "A Machine Learning Implementation for Mental Health Care. Application: Smart Watch for Depression Detection," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 568-574, doi: 10.1109/Confluence51648.2021.9377199

[7] W. H. Organisation, "Depression," World Health Organisation, 30January 2020. [Online]. Available:

[8] A. Menas, "The Widening Mental Health Treatment Gap in Schools,"NEA, 25 February 2019. [Online]. Available:

[9] S. R. Kamite and V. B. Kamble, "Detection of Depression in Social Media via Twitter Using Machine learning Approach," 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), Aurangabad, India, 2020, pp. 122-125, doi:10.1109/ICSIDEMPC49020.2020.9299641

[10] N. A. Asad, M. A. Mahmud Pranto, S. Afreen and M. M. Islam, "Depression Detection by Analyzing Social Media Posts of User," 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems(SPICSCON), Dhaka, Bangladesh, 2019, pp. 13-17, doi: 10.1109/SPICSCON48833.2019.9065101.