TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13844425

Mental Health Prediction Using Machine Learning Algorithms

Vishakha Pathak[a], Kanishka Dwivedi[b] and Mahesh Kumar Tiwari [c]

[a] Student, National PG College, Lucknow, India
[b] Student, National PG College, Lucknow, India
[c] Assistance Professor, Computer Science Department, National PG College, Lucknow, India
vishakhapathak833@gmail.com , kanishkadwivedi22@gmail.com , maheshyogi26@gmail.com

KEYWORD

Mental health; Mental health prediction; Logistic Regression; Random Forest Classifier; Decision Tree Classifier

ABSTRACT

Different mental health issues, such as depression, stress, anxiety, and others, affect life in such a way that it is difficult for personal and professional life to not get severely ruined if not treated on time. The immediate recognition of the condition will ensure proper and best treatment for the patient so that they get the benefit of good health. This research applies Random Forest Classifier, Logistic Regression and Decision Tree Algorithms to evaluate their precision in mental health problem prediction and the stoppage of stress becomes the focus. The evaluation and comparison of their performance are, therefore, useful in bringing out the most reliable way to ensure early intervention for enhanced mental health and support for improved outcomes.

1. Introduction

Mental health refers to the mental state of a person and the general context in which that person is living. Mental illness is usually born from abnormalities in brain chemistry, but one's mental health can also reveal how good one is facing life's troubles and at the same time being stable. [1] To foresee future health issues, it is very important to keep track of the mental health profiles of various groups of people, from working adults, college students to high school youngsters. Mental health is a crucial criterion for tackling primary problems, and the timely detection of these issues can help in the prevention of serious diseases.[3] Stress and sadness, especially, are conventionally believed to afflict persons of all age-groups and economic statuses. Hence, regular check-ups for various parts of the society are the key to stop the progression of the diseases that become worse. [4]

Akin to the former, fixing a person's exact mental condition just by looking at him/her or observing his/her behaviour is a difficult accomplishment of psychological science, and it has not been fully automated. Even though there are instruments for the screening of mental health in almost every place, time and cost may still be a problem for many people.[5] Additionally, a diagnosis-oriented approach will make people reject solutions as a result mental health issues will either go unidentified or ignored.

Corresponding Author: Vishakha Pathak, National Post Graduate College, Lucknow, India
Email: vishakhapathak833@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13844425

Technology development has opened new paths toward the understanding and treatment of mental conditions. Wearables, social media, smartphones, and scanning allow researchers and medical professionals to easily and efficiently collect enormous amounts of data.[6]

Among all techniques, machine learning has found its place in examining data related to mental health in many respects. Similar to how machine learning techniques currently work in mental health data analysis and can improve the treatment outcomes of patients and increase our understanding of mental diseases and how to treat them.[9]

2. Past Works

Thuy Trinh Nguyen et al. looked at review of multimodal machine learning in the detection of mental disorders which is the degree of depression, stress, bipolar disorders. This study shows how the effectiveness of multimodal methods is higher compared to the standalone mode that also brings several main ideas up, and they are the domain-specific and pre-trained embedding that play a big part in learning. The review has unveiled different directions in the field which are along the lines of multi-fusion, co-learning, and the use of deep neural networks.[1]

The work presented by I. Ameer et al. deals with the detection and classification of mental illnesses in social media text, specifically Reddit using the approaches of machine learning, deep learning and transfer learning. Its focus is identifying and assisting in isolating and identifying all the cases of users who require emergency assistance. Assuming the essence of the present-day problem of bipolar disorder, this study is based on the unstructured data of the users where the users are classified into the five most common medically researched – depression, anxiety, bipolar disorder, ADHD, and PTSD.[2]

Stress Detection using Machine Learning Algorithms describes the identification of a stress response by the use of various machine learning methods. The statistical dataset collected is taking into consideration six attributes, namely Electrocardiogram, Electromyogram, Galvanic Skin Response Hand and Foot, Heart Rate, and Respiration to predict one's level of stress. These authors have used different techniques for the classification of the dataset: Decision Tree, Naïve Bayes, and K-Nearest Neighbor; hence, they obtained high accuracy. [8]

Chirantan Ganguly et al discuss the impact of COVID-19 on mental health, the role of machine learning in addressing mental disorders, challenges, and potential solutions. It discusses stigma, access to therapy, data collection technologies, affected populations, ML applications, challenges, and ethical considerations in mental health research and diagnosis using ML models.[4]

The paper by Krishna Shrestha et al, gives a background on the medical condition depression, the role of twitter in prediction, as well as, machine learning. It considers the earlier works that have been published on the use of machine learning in detection of depression and provides recommendation for the subsequent research on the subject matter.[13]

3. Materials and Implementations

This research mainly focuses on finding stress among the individuals with the help of DASS-21 where data was collected from 90 different individuals through Google forms. The data collected was classified using three different methods from supervised learning namely – Logistic Regression and Random Forest Method and Decision Tree.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13844425

3.1. Questionnaires

Age wise, data was collected on the DASS-21 (Depression, Anxiety and Stress Scale) scale. Whereas DASS-21 typically consists of 21 questions (7 per factor) devoted to stress, anxiety and depression in this research the main interest was in consequences of stress. [20]As we limited the scope to 4 direct questions around stress. Participants provided their responses, with the following scale:

0: Did not apply to me at all
1: Applied to me to some degree or time
2: Applied to me to a considerable degree or time
3: Applied to me for most of the time

The responses were collected as numeric values (0 to 3). [20] The adjusted score was determined using a particular method to represent the scaled equivalent of the complete 7-question DASS-21 score because only 4 questions were asked.

$$\text{Adjusted score} = (\text{Sum of 4 questions}/4) \times 7$$

Eq. (A.1)

The four questions asked for determining stress in individuals are listed in table A.1.

Serial No.	Questions asked
1	I tended to overreact to situations.
2	I felt that I was using a lot of nervous energy.
3	I found it difficult to relax.
4	I was intolerant of anything that kept me from getting on with what I was doing.

Table A.1

3.2. Implementation

First of all, dedicatedly built questionnaire was employed in conducting the mental health survey, which included stress and other related topics queries. We collect the responses to our questions, clear the data, and construct this dataset for analysis. To make the data compatible for machine learning algorithms, it went through scaling, validation of missing values, and formatted in the stage of processing.

After partitioning the data for training and test sets, we created models to predict if a student is suffering from mental health issues mainly stress by using classifiers like Random Forest, Logistic Regression etc. To have the best accuracy in the prediction of outcomes such as stress, the performances of these models on unseen data were compared, as this helped us to establish which of these approaches would generate more reliable results.

Formula used for finding the accuracy rate is:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \qquad \text{Eq. (B.1)}$$

In eq. B.1:
TP (True Positive) presents no. of instances where the model correctly predicted high stress when the actual stress level was high

TN (True Negative) presents no. of instances where the model correctly predicted low stress when the actual stress level was low

FP (False Positive) presents no. of instances where the model predicted high stress when the actual stress level was low

FN (False Negative) presents no. of instances where the model predicted low stress when the actual stress level was high

4. Classification Methods

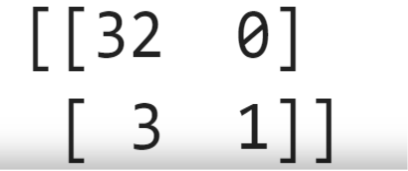
Three supervised learning algorithms are used for classification of the people having stress and people who don’t have stress.

4.1. Logistic Regression

Logistic regression is a class of probabilistic statistical classification models. This method allows for the statistical probability for a given output belonging to two different sets. Simply explained, logistic regression is a model that try to explain the relationship between some binary outcome variable with the dependent variables.

Logistic regression falls within the forms of statistical techniques that are applied in solving problems that involve binary classification. This methodology gives the probability of an outcome falling into one of the two groups. The result is, therefore, that logistic regression attempts to find a model that specifies and differentiates how a set of predictive variables is related to a binary dependent variable.

The confusion matrix created for Logistic Regression is in fig 4.1.A.



```
[[32  0]
 [ 3  1]]
```

Fig 4.1.A

The visualization of fig 4.1.A is mention in fig 4.1.B, offering a detailed assessment of how the model performed in classifying participants' responses.

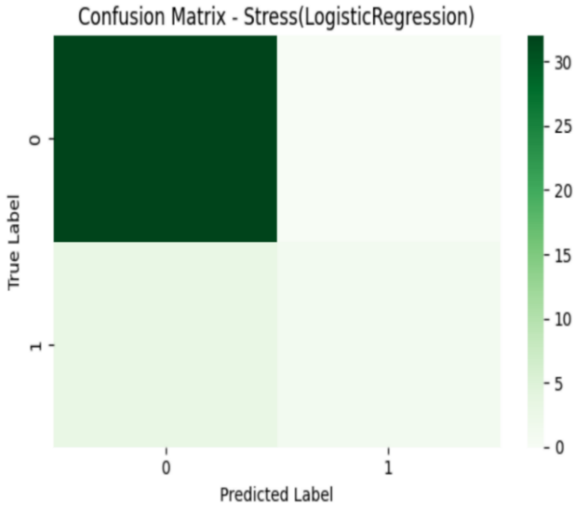


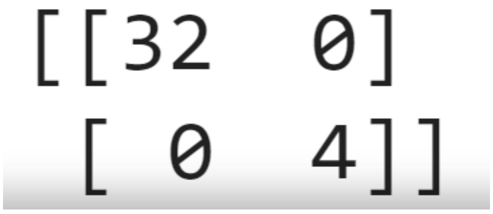
Fig 4.1.B

4.2. Random Forest Classifier

In order to improve prediction and accuracy perseverance, a strong supervised machine learning technique called Random Forest is used. In Random Forest, multiple trees are created by taking a selected portion of data together with a random selection of features. This model generates a group of decision trees and then combines the predictions from each tree to produce an end conclusion that could result in prediction or classification. It is classified as a collection of decisions made by a multitude of trees, which in other terms may be referred to as a "Random Forest". Random Forests are going to be helpful especially in classification tasks when one wants to put an observation into a category and regression tasks when one wants to predict a continuous value.

By using the data collected, the Random Forest method is used in this research to evaluate the stress level. To evaluate the model's performance, a confusion matrix was created, which made it possible to get into a deeper breakdown of the true positives, true negatives, false positives, and false negatives. Such matrix gives a very clear view of how well the model has done in classifying the responses correctly.

The confusion matrix created for the Random Forest Classifier is in fig 4.2.A.



```
[[32  0]
 [ 0  4]]
```

Fig 4.2.A

The visualization of the fig 4.2.A is mention in fig 4.2.B, showcasing the effectiveness of the model in making accurate predictions and providing insights into its classification.

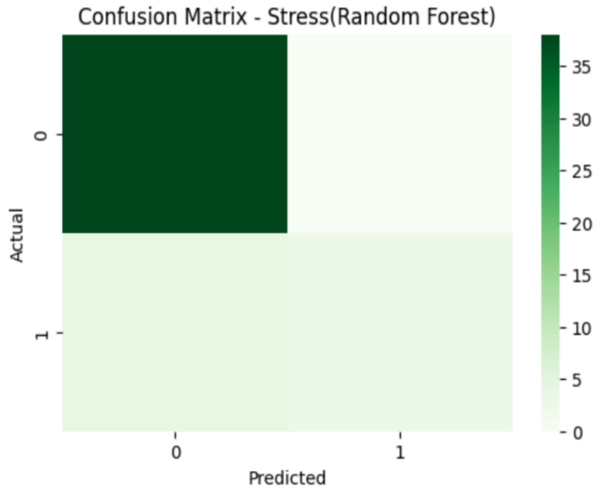


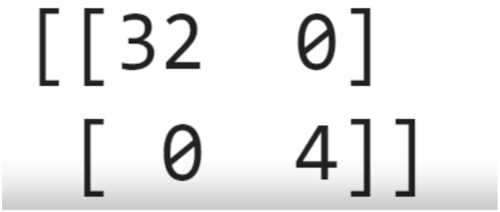
Fig 4.2.B

4.3. Decision Tree Classifier

Decision trees classify data in a manner that is everything from simple, interpretable, and effective presentation of categorical and numerical data. A general strength of decision trees is the possibility to extract descriptive decision rules from the given data, which usually turns them into very valuable insights about how to understand relationships between input features and a target variable.

The decision tree is constructed using a set of data from which the algorithm can learn—using the best splits in the data space to arrive at the most accurate categorization. The aim is to generate a model that, when applied to new data, accurately represents the underlying decision-making process based on the training set.

The confusion matrix created for Decision Tree Classifier is in fig 4.3.A.



```
[[32  0]
 [ 0  4]]
```

Fig 4.3.A

The visualization of the fig 4.3.A is mention in fig 4.3.B, this matrix serves as a critical evaluation tool for understanding how well the model performed in accurately classifying the stress responses.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13844425

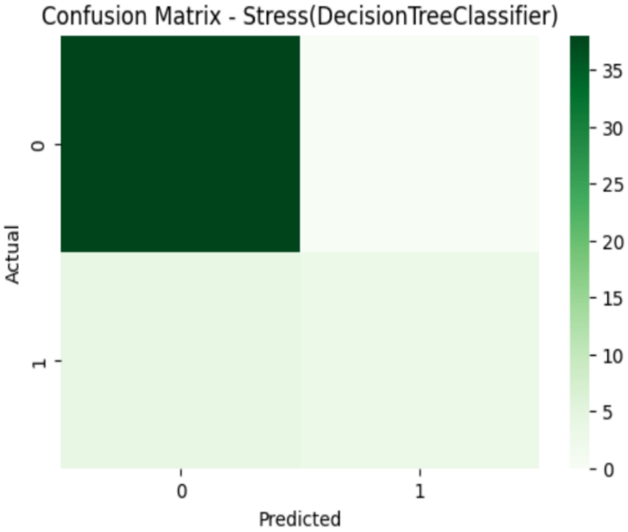


Fig 4.3.B

5. Result

Classification Method	**Accuracy Rate**
Logistic Regression	92%
Random Forest Classifier	100%
Decision Tree Classifier	100%

Table A.1

Random Forest classified stress with 100% accuracy and Decision Tree did the same, so we can infer that both models didn’t make any mistake in classifying each instance in the dataset. So, we can make an inference that these tree-based models were able to recognize the patterns and relationships in stress data.

In this case Logistic Regression had 92% accuracy in classifying stress. Although this is still a very good result, it means it misclassified some, which resulted to 8% error rate. This misclassification might be because of its assumption of linear relationship between independent and dependent variables which might not be good enough to capture the complexity of stress data unlike the flexible nonlinear nature of Random Forest and Decision Tree. This big difference in performance especially the 100% accuracy of tree-based models is an indication that Random Forest and Decision Tree is relatively better in this dataset for stress classification.

6. Challenges

It is complicated to predict problems related to mental health, such as stress, due to imbalanced datasets, situations in which a big proportion of the population suffers from low levels of stress, and factors hard for models to find high. Examples of those models are Random Forest, Decision Trees, and Logistic Regression. The principal problem these will address is that the dataset is imbalanced since the majority of the population can show low levels of stress. Both Random Forest and Decision Trees are resistant to imbalance, but when data becomes extremely

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13844425

imbalanced, there is still a chance it may affect results. Another challenge can be presented by non-linear relationships, which often come up in mental health data. In logistic regression, linearity between features and an outcome is assumed, perhaps leading to a drop in accuracy where the relationships are more complex. On the other hand, Decision Tree and Random Forest are better positioned to handle non-linear data, but they risk overfitting-especially when decision trees get too specific to the training data.

Another domain where interpretability is of great importance is in research on mental health. On one side, Logistic Regression yields an interpretable output with coefficients, but the linearity of the model risks over-simplifying complex data. The intuitive, visually clear paths that Decision Trees produce become less interpretable as complexity increases. Although Random Forests are ensemble in nature, they make decisions that are not as straightforward even though they tend to be more accurate.

Most of the research in mental health is based on sparse data with small sample sizes, further complicating model performance. The Logistic Regression tends to underperform when working on small data; Decision Trees might over fit. Random Forests handle small datasets quite well; however, this can go wrong if the diversity in data is at a minimum. Overall speaking, both Random Forest and Decision Trees outperform Logistic Regression in modelling nonlinear relationships; however, they are suffering from some challenges such as overfitting and sensitivity to noisy data. That means careful model tuning and model evaluation are necessary when making a prediction about mental health.

7. Conclusion

Overall industry wide machine learning is about to change the world. It will take automation to the next level in manufacturing, logistics and customer servicing. For example, diagnoses will be more accurate so healthcare will be more effective and outbreaks will be forecasted better than before. In predicting mental health issues in the future machine learning will revolve around quality and diversity of data for these individual specific factors to become predictors. So real time monitoring and early intervention will become common too. In summary ethical considerations must come first when dealing with this topic area and data privacy. When machine learning is used on sensitive mental health data security issues arise including transparency and informed consent so as not to lose public trust through misuse of data collected from these individuals in totality, all these will lead to machine learning being more accurate, relevant and useful for mental health care.

References

[1]. Multimodal Machine Learning for Mental Disorder Detection: A Scoping Review by Thuy Trinh Nguyen et al.

[2]. Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning by I. Ameer et al

[3]. Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms by Anu Priya et al.

[4]. Mental health impact of COVID19 and machine learning applications in combating mental disorders: a review by Chirantan Ganguly et al

[5]. Machine Learning for Depression Diagnosis using Twitter data by Krishna Shrestha et al.

[6]. Application of Machine Learning Techniques to Predict Depression in social media by M. R. T. et al.

[7]. Using Machine Learning to Detect and Predict Anxiety and Depression from Digital Health Data by Trevor Cohen et al

[8]. Stress Detection using Machine Learning Algorithms by M. R. et al.

[9]. Predicting Mental Health Problems in College Students with Machine Learning Models by J. Wang et al.

[10]. Machine Learning Approaches to Predict Depression, Anxiety, and Stress from Social Media Data by K. Guntuku et al.

[11]. Predicting Anxiety and Depression Disorders in a Large Sample of Psychiatric Patients Using Machine Learning Models by J. Wu, M. Yan, et al.

[12]. Mental Health Care for All: A Call for Action to Address Disparities in Access, Quality, and Outcomes by P. H. Wise et al.

[13]. Understanding Depression with Machine Learning based on Twitter data by Krishna Shrestha et al

[14]. Stress Detection Using Machine Learning Algorithms by V. R. Archana1, B. M. Devaraju.

[15]. Machine Learning in ADHD and Depression Mental Health Diagnosis: A Survey by C. Nash et al.

[16]. HCET: Hierarchical clinical embedding with topic modelling on electronic health records for predicting future depression by Y. Meng, W. Speier, M. Ong, and C. W. Arnold.

[17]. Graham, S.; Depp, C.; Lee, E.E.; Nebeker, C.; Tu, X.; Kim, H.-C.; Jeste, D.V. Artificial Intelligence for Mental Health and Mental Illnesses: An Overview

[18]. Towards Assessing Changes in Degree of Depression through Facebook by H. A. Schwartz et al.

[19]. Lee EE, Torus J, De Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom.

[20]. Sushmita Goswami, Deepak Kumar Chaubey, The Impact of social media on the Spread of Fake News and the Role of Machine Learning in Detection, TEJAS Journal of Technologies and Humanitarian Science, ISSN-2583-5599, Vol.02, I.01(2023)

[21]. Neha Singha and Yogendra Pratap Singh, Consumer Sentiment Analysis Using Deep Learning, TEJAS Journal of Technologies and Humanitarian Science, ISSN-2583-5599,Vol.02, I.02(2023)