PREDICTION OF MENTAL HEALTH DISORDERS IN TEENAGERS USING MACHINE LEARNING ALGORITHMS AND SUGGESTING SOLUTIONS FOR THE PROBLEMS

CS 584 PROJECT REPORT

Kajol Tanesh Shah
Hawk Id: A20496724

Surya Thatee
Hawk Id :A20473893

Illinois Institute Of Technology

December 2021

ABSTRACT

Major career decisions of a human are to be decided between the ages 13-19 and also the physical and pyschosocial growth happens at this age. In some cases, these are affected by stress, anxiety and depression of that person. In this project, we developed a model to predict stress, anxiety and depression for teenagers using machine learning algorithms. For this, data was collected with an online version of Depression Anxiety Stress Scale(DASS 42).This data contains many attributes like response values, response time, country, age, education, etc. In our project, we have used five machine learning algorithms to predict the five-levels of severity for stress, anxiety and depression.

1 Introduction

Mental health refers to our emotional, psychological, and social well-being. Our thinking abilities, feelings and actions are affected by our mental health. Research says that half of the mental illnesses begin by age 14. So, it is critical to intervene early to minimize its effects on the development, education, employment and health of a person. It is important to be aware about mental health and busting the stigma around it. Fears and worries are typical in a person but these fears and worries maybe because of stress, anxiety, and depression. These three disorders seem to be similar but each has their unique causes and symptoms.

Frustrated, anxious, irritable are the some of the symptoms of stress. Dealing with stress depends on the persons ability to handle situations like social, economic circumstances etc.In long term, if stress is not handled it may lead to physical or mental illness.

Loss of interest, changes in appetite, loss of energy, feeling worthless are some of the symptoms of depression. Depression in most severe cases can also be treated with antidepressants, psychotherapy etc. Depression often gets worse if it isn't treated, resulting in emotional, behavioral and health problems that affect every area of your life.

In our daily life, feeling anxious is very common. Sometimes, people with anxiety disorders frequently have intense, excessive and persistent worry and fear about everyday situations. Anxiety leads to panic attacks which in cases may become severe leading to heart attacks, trembling, sensations of shortness of breath etc. High level anxiety can be treated with medications but it can only relive you of some of the symptoms. Anxiety is something that should be handled with self-care like getting control of mind and body with yoga, exercise or meditations, etc.

Some of the statistics considered from [] related to mental illnesses are 1) 50 of all lifetime mental illness begins by age 14, and 75 by age 24. 2) Suicide is the 2nd leading cause of death among people aged 10-34. 3) 1 in 5 U.S. adults experience mental illness each year. 4) 1 in 20 U.S. adults experience serious mental illness each year. 5) 1 in 6 U.S. youth aged 6-17 experience a mental health disorder each year. In our project we focused primarily on teenagers with symptoms of stress, anxiety and depression.The data collected was filtered down to age group of 13-19.All the data used for machine learning algorithms are trained to predicts the mental health conditions in this age group.

For models, the data is collected using standard diagnosis criterion for depression using the Patient Health Questionnaire (PHQ).This questionnaire contains 42 questions where each of stress, anxiety and depression are allocated 14 questions each and based on the responses of these questions, we measure the severity of the mental disorders using Depression, Anxiety and Stress Scales DASS(42).

Later sections in the report contains Related Studies, Materials, Methods, Results, Problems.

2 Related Studies

Previous research on predicting stress, anxiety and depression used machine learning algorithms such as Logistic Regression, Naive Bayes, Random Forest Tree and Support Vector Machine. In most scenarios, model trained gave promising results for most of the algorithms.

Some researchers used social media data to explore interaction of various aspects of mental health, including depression, anxiety, suicidal ideas, mood and stress.

Researchers also developed separate models for each stress, anxiety and depression. In such cases, all the data used and model created have robust and lot of conditions or symptoms been considered. Which leads to a more accurate predictions.

Natural Language Processing (NLP) techniques and machine learning approaches are used to detect depression in Reddit social media forum. They found large number of lexicon of terms that are similar among depressed accounts.They mentioned Support Vector Machine classifier detected depression at an accuracy of 80 percent.

To predict mental health conditions, multi-modal cues are used from the social media posts. Multi-modal approach consists of jointly analyzing language, visual, and metadata cues and their relation to mental health. They classified by using experiments where they discriminated between healthy users and and users affected by mental health illness.

3 Materials

Depression, Anxiety and Stress Scale questionnaire (DASS 42) was used to train the models. Dataset was from Kaggle which has data collected from online version of Depression Anxiety Stress Scale(DASS 42) questionnaire. It contains responses of around 40000 people with features including answers to the questions, response time taken to answer the questions, and other characteristics of the person like country, age, personality, education, etc.

This dataset has response values in the range 1 – 4 but ideally DASS(42) considers response values in the range 0 – 3. So, we changed the scales in the dataset from 1 – 4 to 0 – 3 . As our focus is on teenagers, we considered the age group 13 – 19 and filtered the dataset on this age group. Dataset doesn't contain outcome columns, so based on the DASS(42) template and scales we created the outcome columns for Stress, Anxiety, Depression by adding the respective response values. As per DASS(42), response time for each question should not be more. So, in our model we didn't consider the questions with response time more than 25000 milliseconds.

Dass(42) contains total of 42 questions which has to be answered. Each stress, anxiety, depression we allocated with 14 questions. Below are few questions from each section. Each question is provided with 4 choices, which are as follows in the fig.

	Anxiety	Depression	Stress
1.	Dryness of Mouth	Couldn't Experience the positive feeling	Found hard to wind down
2.	Difficulty in Breathing	Difficult to work up the initiative to do things	Overreact to situations
3.	Experience Trembling	Nothing to look forward	A lot of nervous energy
4.	Worried about panic and make a fool of themselves	Felt down-hearted and Blue	Getting Agitated
5.	Close to Panic	Unable to become enthusiastic	Difficult to Relax
6.	Aware of the action of the heart in the absence of physical exertion	Felt wasn't worth much as a person	Intolerant to getting what I was doing
7.	Felt scared without any good reason	Felt life was meaningless	Touchy

	Depression	**Anxiety**	**Stress**
Normal	0-9	0-7	0-14
Mild	10-13	8-9	15-18
Moderate	14-20	10-14	19-25
Severe	21-27	15-19	26-33
Extremely Severe	28+	20+	34+

The rating scale is as follows:

0 Did not apply to me at all
1 Applied to me to some degree, or some of the time
2 Applied to me to a considerable degree or a good part of time
3 Applied to me very much or most of the time

In the above second figure the DASS(42) standard scales used to measure the severity levels of the three mental disorders are mentioned.

4 Methods

We have used Python Programming language to apply machine learning algorithms to the dataset. With required libraries being loaded to train models, the dataset was prepared for performing the approach. The dataset contains total of 175 columns where only 43 of them are important for predicting the outcomes. This dataset doesn't contain any outcome column, so the outcome columns were created as per the DASS standards and scales and were added to the data. Outcome columns are Stress, Anxiety and Depression. Each output variable column depends upon 14 features out of the 42 features and the age feature. The output variable value lies in the range of the scales of Stress, Anxiety and Depression (Fig)

Classification algorithms of supervised machine learning are used to predict the outcomes. In classification, the algorithm is trained on the training dataset and based on that training, it categorizes the data into different classes.

The algorithms used to train the model are 1) Logistic Regression 2) Random Forest 3) Decision Tree 4) Support Vector Machine 5) Naive Bayes

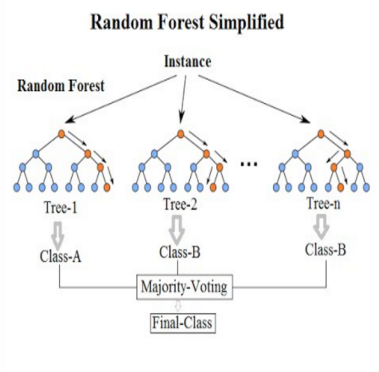
4.1 Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. It is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

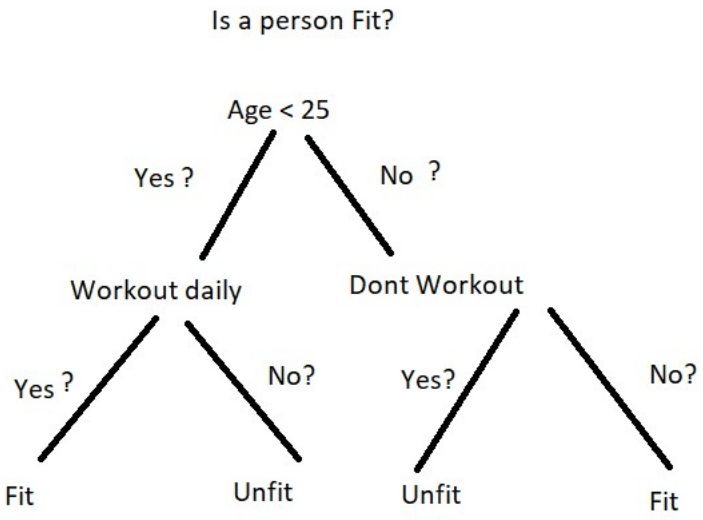
4.2 Random Forest

Random forest builds multiple Decision trees and merges them together to get a more accurate and stable prediction. It collects the points from different decision trees and decides the final class of test objects.



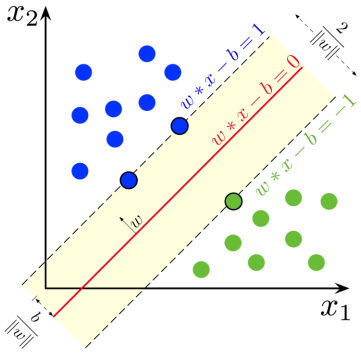
4.3 Decision Tree

In Supervised Machine Learning, Decision tree is a method where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.



4.4 Support Vector Machine

Support Vector Machine can be used for both Classification and Regression problems. It aims to create a hyper-plane in a N-dimensional (N is number of features) space that classifies the data points. Fig- Maximum-margin hyper-plane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.



4.5 Naive Bayes

For Naive Bayes, Bayes theorem is used for assigning class labels to Problem instances.

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Approach: We used two approaches as in our case, we are predicting multiple outputs – Stress, Anxiety and Depression

In our first approach, we used all the selected 43 features for training the model to predict 3 outputs using MultiOutputRegressor. Here, we are predicting the values which are in the range of the DASS(42) scales which we have mentioned in the previous slide. Then, based on the values predicted, we are classifying them in the given categories.

In our second approach, we classified the 3 outcome variables beforehand into categories based on their scale values and then trained the model. Here, instead of training 1 model for all the 3 outputs, we created 1 model per output variable i.e. 3 models for 3 output variables. For stress model, out of the 43 variables, we trained our model only on 15 features on which the Stress output was dependent. we used a similar approach for training other two models too.

5 Results

5.1 Approach 1

Trained models are evaluated on the test sets for all the models. Since we considered all the features for this approach, while evaluating the test set, the model takes into consideration all the three output variables which leads to less accuracy for the Linear regression, Random Forest, Decision tree.

Model	Accuracy
Linear Regression	51%
Random Forest	28.13%
Decision Tree	20.25%
Support Vector Machine	99.97%

As we can observe from above fig,In this approach except for SVM none of the other models gave proper accuracies. So we followed a 2nd approach which is as mentioned below.

5.2 Approach 2

In this approach, we trained 3 models for 3 output variables i.e. 1 model per output variable. Here, only 14 question features and 1 age feature were considered for training a single model. So, the respective models for stress, anxiety and depression were trained only on 15 features which were related to those models. This approach resulted in better accuracies than the previous approach.

Classifier	Mental Illness	Accuracy	Precision	Recall	F1 score
Logistic Regression	Extremely severe Anxiety		1	1	1
	Mild Anxiety		1	0.99	1
	Moderate Anxiety	99.96%	1	1	1
	Normal Anxiety		1	1	1
	Severe Anxiety		1	1	1
Random Forest	Extremely severe Anxiety		0.96	0.97	0.97
	Mild Anxiety		0.69	0.2	0.31
	Moderate Anxiety	88.06%	0.74	0.88	0.8
	Normal Anxiety		0.85	0.98	0.91
	Severe Anxiety		0.84	0.78	0.81
Decision Tree	Extremely severe Anxiety		0.92	0.91	0.91
	Mild Anxiety		0.38	0.37	0.38
	Moderate Anxiety	75.18%	0.55	0.57	0.56
	Normal Anxiety		0.84	0.86	0.85
	Severe Anxiety		0.52	0.5	0.51
Naive Bayes	Extremely severe Anxiety		1	0.86	0.93
	Mild Anxiety		0.52	0.34	0.41
	Moderate Anxiety	82.09%	0.69	0.83	0.75
	Normal Anxiety		0.91	0.86	0.88
	Severe Anxiety		0.61	0.85	0.71
Support Vector Machine	Extremely severe Anxiety		1	1	1
	Mild Anxiety		1	1	1
	Moderate Anxiety	100%	1	1	1
	Normal Anxiety		1	1	1
	Severe Anxiety		1	1	1

The above figure contains the metrics achieved for the Anxiety model using different machine learning algorithms.

Classifier	Mental Illness	Accuracy	Precision	Recall	F1 score
Logistic Regression	Extremely severe Stress		1	1	1
	Mild Stress		1	1	1
	Moderate Stress	99.96%	1	1	1
	Normal Stress		1	1	1
	Severe Stress		1	1	1
Random Forest	Extremely severe Stress		0.97	0.92	0.95
	Mild Stress		0.86	0.65	0.74
	Moderate Stress	90.09%	0.85	0.92	0.88
	Normal Stress		0.93	0.97	0.95
	Severe Stress		0.89	0.93	0.91
Decision Tree	Extremely severe Stress		0.83	0.85	0.84
	Mild Stress		0.5	0.48	0.49
	Moderate Stress	73.74%	0.69	0.67	0.68
	Normal Stress		0.84	0.88	0.86
	Severe Stress		0.72	0.71	0.72
Naive Bayes	Extremely severe Stress		0.96	0.86	0.91
	Mild Stress		0.75	0.85	0.8
	Moderate Stress	89.47%	0.87	0.94	0.9
	Normal Stress		1	0.88	0.93
	Severe Stress		0.87	0.92	0.89
Support Vector Machine	Extremely severe Stress		1	1	1
	Mild Stress		1	1	1
	Moderate Stress	100%	1	1	1
	Normal Stress		1	1	1
	Severe Stress		1	1	1

The above figure contains the metrics achieved for the Stress model using different machine learning algorithms.

Classifier	Mental Illness	Accuracy	Precision	Recall	F1 score
Logistic Regression	Extremely severe Depression	99.27%	1	1	1
	Mild Depression		0.97	1	0.98
	Moderate Depression		1	0.68	0.81
	Normal Depression		1	1	1
	Severe Depression		0.96	1	0.98
Random Forest	Extremely severe Depression	93.16%	0.98	0.99	0.98
	Mild Depression		0.8	0.76	0.78
	Moderate Depression		0.56	0.08	0.14
	Normal Depression		0.93	0.98	0.95
	Severe Depression		0.81	0.84	0.82
Decision Tree	Extremely severe Depression	85.49%	0.96	0.95	0.96
	Mild Depression		0.56	0.56	0.56
	Moderate Depression		0.16	0.16	0.16
	Normal Depression		0.9	0.89	0.89
	Severe Depression		0.61	0.64	0.63
Naive Bayes	Extremely severe Depression	85.42%	1	0.86	0.92
	Mild Depression		0.67	0.87	0.76
	Moderate Depression		0	0	0
	Normal Depression		1	0.87	0.93
	Severe Depression		0.52	0.98	0.68
Support Vector Machine	Extremely severe Depression	100%	1	1	1
	Mild Depression		1	1	1
	Moderate Depression		1	1	1
	Normal Depression		1	1	1
	Severe Depression		1	1	1

The above figure contains the metrics achieved for the Depression model using different machine learning algorithms.

6 Problems Faced

The dataset on Kaggle did not have the outcome columns. So, by understanding the DASS(42) questionnaire, template and scales, we generated the three outcome columns for Depression, Anxiety and Stress

In the traditional machine learning approaches, we usually have only one outcome column but here, we had to predict 3. So, we figured out 2 approaches to solve this problem: one with the Regression approach using MultioutputRegressor for predicting multiple outputs at the same time and the other, training 3 different models for the 3 outputs using Classification approach.

7 Conclusion and Future Scope

7.1 Conclusion

For this model, we took response values of the users to train the model and predict the mental disorder on a scale provided by DASS(42). The five level severity for each Stress, Anxiety and Depression are determined with the ML models. The second approach gave promising accuracies in predicting the outcomes for all the models. We achieved better accuracies in the second approach as we performed feature selection.

7.2 Future Scope

We can create an application using this model. The user will have to answer all the 42 questions and then, based on that, the mental disorders will be predicted.

We have created this model for teenagers but we can also create a similar type of model for adults too. Along with the prediction of the mental disorders, we can also provide the user with some basic solutions for the treatment of the disorders.

With data on the past diagnosed mental illnesses, we can also suggest required medications for the predicted disorders.

8 Reference

1)https://www.sciencedirect.com/science/article/pii/S1877050920309091

2)https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/2017$_{N}ational_{M}ental$

3)https://www.frontiersin.org/articles/10.3389/fdata.2020.00015/full

4)https://dl.acm.org/doi/fullHtml/10.1145/3398069

5)https://academic.oup.com/tbm/article-abstract/9/6/1197/5369573?redirectedFrom=fullt

6)https://www.sciencedaily.com/releases/2021/02/210208114230.html

7)https://aclanthology.org/2020.lrec-1.772/

8)https://ieeexplore.ieee.org/abstract/document/8681445

9)https://www.kaggle.com/lucasgreenwell/depression-anxiety-stress-scales-responses?select=codebook.txt
10)Dataset:-https://www.sciencedirect.com/science/article/pii/S1877050920309091

11)https://dl.acm.org/doi/fullHtml/10.1145/3308558.3313557

12)http://www2.psy.unsw.edu.au/dass/

9 GitHub Link

GitHub Link (Surya Thatee) -https://github.com/suryat9988/ML-Project.git

GitHub Link (Kajol Tanesh Shah) -https://github.com/kajolshah310/Machine-learning-project