PREDICTING DEPRESSION USING MACHINE LEARNING

MINI PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY ENGINEERING

by

Raghav Maheshwari
Enrollment No: 00111503119

Meghdoot Panda
Enrollment No: 00211503119

Harshit Mathur
Enrollment No: 01711503119

Guided by

Mr. Ravi Arora
Asst. Professor

DEPARTMENT OF INFORMATION TECHNOLOGY ENGINEERING
BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING
(AFFILIATED TO GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY, DELHI)
DELHI – 110063

MAY 2022

CANDIDATE'S DECLARATION

It is hereby certified that the work which is being presented in the B. Tech Major Project Report entitled **"PREDICTING DEPRESSION USING MACHINE LEARNING"** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** and submitted in the **Department of Information Technology Engineering** of **BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING, New Delhi (Affiliated to Guru Gobind Singh Indraprastha University, Delhi)** is an authentic record of our own work carried out during a period from **January 2022 to May 2022** under the guidance of **Mr. Ravi Arora, Asst. Professor.**

The matter presented in the B. Tech Major Project Report has not been submitted by me for the award of any other degree of this or any other Institute.

(Raghav Maheshwari)	**(Meghdoot Panda)**	**(Harshit Mathur)**
(En. No: 00111503119)	**(En. No: 00211503119)**	**(En. No: 01711503119)**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge. He/She/They are permitted to appear in the External Major Project Examination

Mr. Ravi Arora
Asst. professor

Dr. Prakhar Priyadarshi
Head, IT Dept.

The B. Tech Minor Project Viva-Voce Examination of **Name of the Student (Enrollment No: XXX),** has been held on **…….**

(Signature of External Examiner)

Mr. Ravi Arora **Ms. Anshu Khurana**

ABSTRACT

Millions of people globally suffer from depression, and it is a debilitating condition. At best it can be difficult for people to live their lives normally and happily, and at worst it leads to death by suicide. Primary care doctors are overwhelmingly finding that they are faced with the need to treat mental health conditions such as depression without any training of how to handle such cases.

There is evidence that an integrated approach where physicians regularly screen patients for mental health disorders and work together with psychologists and other mental health professionals to treat patients leads to reduced costs and better patient outcomes. However, this approach can require a lot of buy-in from many individuals, require extra training, and is often not logistically feasible.
Using data, machine learning was applied to predict patients who may have depression based on information that could typically be surveyed. These predictions could be used to put patients in touch with experienced mental health professionals sooner and easier.

The results show that 86% of those who have depression and 89% of those who don't have depression can be correctly identified. Though more work needs to be done to create a more accurate model, this shows proof of concept that this is a realistic prediction task. Better results could be yielded by adding more patient information to the data or testing more types of models.

ACKNOWLEDGEMENT

We express our deep gratitude to **Mr. Ravi Arora**, Assistant Professor, Department of Information Technology Engineering for his valuable guidance and suggestion throughout our project work. We are thankful to **Ms. Anshu Khurana**, for her valuable guidance.

We would like to extend my sincere thanks to **Dr. Prakhar Priyadarshi, Head of the Department, IT** for his time to time suggestions to complete my project work. I am also thankful to **Dr. Dharmender Saini, Principal** for providing me the facilities to carry out my project work.

(Raghav Maheshwari)
(En. No: 00111503119)

(Meghdoot Panda)
(En. No: 00211503119)

(Harshit Mathur)
(En. No:01711503119)

TABLE OF CONTENTS

LIST OF FIGURES:

LIST OF SYMBOLS:

$\prod$	The Cartesian product can be generalized to the n-ary Cartesian product over n sets X1, ..., Xn: If tuples are defined as nested ordered pairs, it can be identified to (X1 × ... × Xn-1) × Xn.
$\sum$	Capital-sigma notation represents summation of many similar terms.
∂	Partial derivative of a function of several variables is its derivative with respect to one of those variables, with the others held constant.
log	The logarithm is the inverse function to exponentiation.
e	The number ***e***, also known as Euler's number, is a mathematical constant approximately equal to 2.71828.

CHAPTER-1

1.1 INTRODUCTION

A person's mental well-being is his or her mental condition, as well as an overview of his or her general environment. Brain chemistry abnormalities are the cause of mental illness. An individual's mental health serves as a barometer for properly addressing his or her diseases. To predict any health-related irregularities, it is critical to keep track of diverse groups' mental health profiles. The community is made up of working professionals, college students, and high school students. There is a widespread belief that stress, and sadness affect people of all ages and backgrounds. To avoid serious illness, it is necessary to identify the mental health of different categories at different times. In the next years, healthcare providers will be required to consider a patient's mental health profile to deliver better medication and aid in a speedier recovery.

Anxiety and depression are serious public health issues that affect people all over the world. They affect people of all ages, from children to the elderly, including both men and women. Anxiety and depression disorders have a wide range of effects on health and well-being. They are responsible for a variety of somatic symptoms such as gastritis, acid reflux, palpitation, insomnia or hypersomnia, tremor, significant weight loss or gain, and various psychosocial manifestations such as depressed mood, social withdrawal, decreased workplace productivity, suicidal ideation, or attempt, and lack of concentration. Depression and anxiety are significant risk factors for a variety of other lifestyle disorders, such as ischemic heart disease, hypertension, diabetes, unintentional accidents, and deliberate. Suicidal ideation and depression are intimately linked, and depression can lead to suicide. There are different communicable diseases, such as tuberculosis and HIV, that harm them. Depression and anxiety sufferers are frequently stigmatized by society and socially isolated by their families. In educational institutions and workplaces, they may underperform. As a result, people are losing access to economic and social possibilities, with results in a low quality of life. Economic strain is a massive and often unquantifiable manifestation that contributes to a vicious cycle of poverty and bad health. The majority of those affected are low- and middle-income families.

Technological advancements such as smartphones, social media, neuroimaging, and wearables have enabled researchers of mental health and doctors to gather a tremendous amount of information at a rapid rate. Machine learning has developed as a reliable tool for analysing these data. Machine Learning is the application of advanced probabilistic and statistical techniques to create computers that can learn from data on their own. This allows data patterns to be more easily and correctly discovered, as well as more accurate predictions from data sources.

1.2 MOTIVATION

Ever since pandemic hit India, followed by an unprecedented lockdown, stress levels have been on the rise with 43 percent Indians suffering from depression, according to a study. Conducted by GOQii a smart tech enabled preventive healthcare platform, the study surveyed over 10000 Indians to understand how they have been coping with the new normal. According to study, 26 percent respondents were suffering from mild depression, 11 percent were feeling moderately depressed, and 6 percent were having severe symptoms of depression.

The last 2 years have been unexpected. The situation has taken a major toll on the mental health of citizens. With the series of lockdowns, anxiety, job cuts, health scares and overall volatile environment, stress levels are on an all-time high. "Copious amounts of stress can lead to depression. With the current lockdowns and lifestyles drastically changing, we have seen that 43% of Indians are currently plagued with depression", the study said.

The unprecedented experience of 'home quarantine' under lockdown with the uncertainty of academic and professional career has multifaceted impacts on the mental health of the people. For example, a Canadian study focusing on the effects of quarantine after the severe acute respiratory syndrome (SARS) epidemic found an association between longer duration of quarantine with a high prevalence of anxiety and depression among people .The ongoing COVID-19 pandemic is creating a psycho-emotional chaotic situation as countries have been reporting a sharp rise of mental health problems, including anxiety, depression, stress, sleep disorder as well as fear, among its citizens, that eventually increased the substance use and sometimes suicidal behaviour. Researchers in China observed that the greater exposure to 'misinformation' through social media are more likely contributing to the development of anxiety, depression, and other mental health problems among its population of different socioeconomic background. Studies before the COVID-19 pandemic also suggested an inverse relationship between media exposure and mental health. On the contrary, a study in South Korea during the Middle East respiratory syndrome (MERS) reported a positive relationship between risk perception and media exposure.

Given the unexpected circumstances, it is crucial to explore the psycho-social experience of people in India, especially during the COVID-19 pandemic.

1.3 OBJECTIVE

1. The goal of this project is to create a model that can be used to predict depression in a person based on a series of questions.

2. The subjects who are predicted to have depression could potentially be referred straight to mental health professionals in their area or who accept their health care coverage.

1.4 SUMMARY OF THE REPORT

Chapter 1 discusses about the Introduction to the project, our motivation behind the project and the objectives we have for this project. The Introduction talks about the issues with mental health faced in the current scenario and how we plan on tackling them using latest technologies like machine learning. Our motivation has been the drastic change in the mental health brought upon the population of India due to the COVID-19 pandemic. With the Objectives we discuss our take on tackling the mental health crisis as well as a future aspect for this project.

In Chapter 2, we share details about the project. We talk about the goal, the resources and our approach. The goal is to predict depression in a person using machine learning, specifically logistic regression.

In the third chapter, we discuss the results of our implementation of logistic regression and discuss the influence of various features on the final prediction.

Finally, we conclude with the learnings and the future aspects of this project and how we can move forth with this topic in the final chapter of this report.

Chapter 2

2.1 About the Project

The goal of this project is to gather data about people to predict depression.

Having standard services where patients are constantly screened for mental health disorders and treatment is tightly integrated with teams of physicians and psychological professionals can be expensive, requires a lot of training, requires participation from many individual doctors that may feel too overwhelmed, and may also not be possible in certain areas due to various logistical factors. Using machine learning and data, the goal is to predict who may have depression in a way that requires very little human participation from doctors and has lower time and money costs associated. The patients who are predicted to have depression could potentially be referred straight to mental health professionals in their area or who accept their health care coverage. The patient's file could also be flagged to alert the medical staff the next time they have any kind of physician appointment to prompt doctors to start the conversation with patients. At the very least information and resources could be sent to patients directly to encourage them to act on their own behalf.

2.2 Dataset description:

The data for this project is taken from Kaggle, which is a hub for free datasets. It is home to the world's largest collaborative data community, which is free and open to the public. The dataset consists of 32 columns and 158 rows. The columns were mixed with ordinal as well as nominal data based on which we developed our prediction model.

2.3 APPROACH

2.3.1 Dataset generation:

- Dataset was obtained from Kaggle.
- The dataset has 158 entries including about 158 individuals and 32 columns.

Fig 2.1(a) - Data before pre-processing

Out[4]:

	Timestamp	Email address	Name	Gender	Are you above 30 years of age?	Employment Status	City	How are you feeling today?	eating and sleeping	(If sad)have you been in the same mental state for the past few days?	...	Having trouble concentrating on things, such as reading the newspaper or watching television, or studying?	Do you feel bad about yourself — or that you are a failure or have let yourself or your family down?
0	09-12-2021 23:54	riyaaditi2@gmail.com	Aditi Harsh	Female	No	Student	Tier 3 (Other cities/towns)	Fine	Yes	No	...	No	No
1	10-12-2021 08:40	imcrazyashutosh@gmail.com	Ashutosh Kumar	Male	No	Student	Tier 2 (Capital cities Eg. Lucknow)	Fine	No	Yes	...	Yes	Yes
2	10-12-2021 21:48	atharv23srivastava@gmail.com	Atharv srivastava	Male	No	Student	Tier 1 (Delhi, Mumbai, Bangalore, Chennai, Kol...	Fine	No	Maybe	...	Maybe	Maybe
3	10-12-2021 21:50	ritulricha22@gmail.com	Rimi	Female	No	Student	Tier 3 (Other cities/towns)	Fine	No	Maybe	...	Yes	Yes
4	10-12-2021 21:55	nisha18054@gmail.com	Jaya singh	Female	No	Student	Tier 2 (Capital cities Eg. Lucknow)	Good	Yes	No	...	No	No

5 rows × 32 columns

Fig 2.1(b)

Out[4]:

	How are you feeling today?	eating and sleeping	(If sad)have you been in the same mental state for the past few days?	...	Having trouble concentrating on things, such as reading the newspaper or watching television, or studying?	Do you feel bad about yourself — or that you are a failure or have let yourself or your family down?	How many hours do you spend per day on watching mobile phone, laptop, computer, television, etc.?	If sad, how likely are you to take an appointment with a psychologist or a counsellor for your current mental state?	Has the COVID-19 pandemic affected your mental well being?	How often do you get offended or angry or start crying ?	How likely do you feel yourself vulnerable or lonely?	How comfortable are you in talking about your mental health?	Prediction	Prediction_status
3	Fine	Yes	No	...	No	No	NaN	NaN	NaN	NaN	NaN	NaN	36	Yes
2	Fine	No	Yes	...	Yes	Yes	NaN	NaN	NaN	NaN	NaN	NaN	33	No
1	Fine	No	Maybe	...	Maybe	Maybe	More than 10 hours	1.0	Yes	Sometimes	2.0	NaN	34	No
3	Fine	No	Maybe	...	Yes	Yes	2-5 hours	1.0	Yes	Often	4.0	NaN	35	Yes
2	Good	Yes	No	...	No	No	5-10 hours	1.0	Not sure	Sometimes	4.0	NaN	33	No

2.3.2 Dataset preprocessing and cleaning to get relevant features:

Here the corrupt or inaccurate record from dataset were detected and then corrected.

- There were few NaN cells where we filled with relevant data.
- Irrelevant columns were removed from the dataset.
- Transposing the column to obtain the desired dataset format.
- All the data was converted to binary format. i.e. 0 or 1.

Fig 2.2(a) – Data after processing and cleaning

In [41]:
```
data
```

Out[41]:

	Gender	Are you above 30 years of age?	How are you feeling today?	Is your sadness momentarily or has it been constant for a long time?	At what time of the day are you extremely low?	How frequently have you had little pleasure or interest in the activities you usually enjoy?	How confident you have been feeling in your capabilities recently.	Describe how 'supported' you feel by others around you – your friends, family, or otherwise.	How frequently have you been doing things that mean something to you or your life?	How easy is it for you to take medical leave for a mental health condition?	...	ther_Yes	conc_Maybe	conc_No	conc_Ye
0	0	0	1	1	2	3	4	0	2	2	...	0	0	1	
1	1	0	1	1	0	1	2	2	2	2	...	0	0	0	
2	1	0	1	2	2	0	2	2	3	0	...	0	1	0	
3	0	0	1	2	2	1	2	1	2	2	...	0	0	0	
4	0	0	0	0	1	3	3	1	1	2	...	0	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
153	1	1	0	3	2	2	5	1	1	2	...	0	0	0	
154	1	1	0	2	2	2	5	0	1	1	...	0	0	1	
155	2	0	3	0	0	0	1	3	0	3	...	1	0	0	
156	1	0	1	0	2	0	5	3	3	0	...	0	1	0	
157	0	1	1	1	0	1	4	0	1	1	...	0	0	1	

158 rows × 60 columns

Fig 2.2(b):

In [41]:
```
data
```

Out[41]:

	How confident you have been feeling in your capabilities recently.	Describe how 'supported' you feel by others around you – your friends, family, or otherwise.	How frequently have you been doing things that mean something to you or your life?	How easy is it for you to take medical leave for a mental health condition?	...	ther_Yes	conc_Maybe	conc_No	conc_Yes	fbad_Maybe	fbad_No	fbad_Yes	cov_No	cov_Not sure	cov_Yes
3	4	0	2	2	...	0	0	1	0	0	1	0	0	0	1
1	2	2	2	2	...	0	0	0	1	0	0	1	0	0	1
0	2	2	3	0	...	0	1	0	0	1	0	0	0	0	1
1	2	1	2	2	...	0	0	0	1	0	0	1	0	0	1
3	3	1	1	2	...	0	0	1	0	0	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2	5	1	1	2	...	0	0	0	1	0	0	1	0	0	1
2	5	0	1	1	...	0	0	1	0	0	1	0	0	0	1
0	1	3	0	3	...	1	0	0	1	0	1	0	0	0	1
0	5	3	3	0	...	0	1	0	0	1	0	0	0	1	0
1	4	0	1	1	...	0	0	1	0	1	0	0	0	1	0

(6)

2.3.3 Modelling:

- In this project we used logistic regression as the target variable is categorical

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.
- The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.
- Binary Logistic Regression Model − The simplest form of logistic regression is binary or binomial logistic regression in which the target or dependent variable can have only 2 possible types either 1 or 0.
- Multinomial Logistic Regression Model − Another useful form of logistic regression is multinomial logistic regression in which the target or dependent variable can have 3 or more possible *unordered* types i.e. the types having no quantitative significance.

2.3.4 Maths behind Logistic Regression

We could start by assuming p(x) be the linear function. However, the problem is that p is the probability that should vary from 0 to 1 whereas p(x) is an unbounded linear equation. To address this problem, let us assume, log p(x) be a linear function of x and further, to bound it between a range of (0,1), we will use logit transformation. Therefore, we will consider log p(x)/(1-p(x)). Next, we will make this function to be linear:

$$\log \frac{p(x)}{1-p(x)} = \alpha_0 + \alpha . x$$

After solving for p(x):

$$p(x) = \frac{e^{\alpha_0+\alpha}}{e^{\alpha_0+\alpha}+1}$$

To make the logistic regression a linear classifier, we could choose a certain threshold, e.g. 0.5. Now, the misclassification rate can be minimized if we predict y=1 when $p \geq 0.5$ and y=0 when $p<0.5$. Here, 1 and 0 are the classes.
Since Logistic regression predicts probabilities, we can fit it using likelihood. Therefore, for each training data point x, the predicted class is y. Probability of y is either p if y=1 or 1-p if y=0. Now, the likelihood can be written as:

$$L(\alpha_0, \alpha) = \prod_{I=1}^{n} p(x_i)^{y_i}(1-p(x_i)^{1-y_i}$$

The multiplication can be transformed into a sum by taking the log:

$$l(\alpha_0, \alpha) = \sum_{i=0}^{n} y_i \log p(x_i) + (1 - y_i) log 1 - p(x_i)$$

$$= \sum_{i=0}^{n} \log 1 - p(x_i) + \sum_{i=0}^{n} y_i log \frac{p(x_i)}{1 - p(x_i)}$$

Further, after putting the value of p(x):

$$l(\alpha_0, \alpha) = \sum_{i=0}^{n} - \log 1 + e^{\alpha_0 + \alpha} + \sum_{i=0}^{n} y_i(\alpha_0 + \alpha . x_i)$$

The next step is to take a maximum of the above likelihood function because in the case of logistic regression gradient ascent is implemented (opposite of gradient descent).

Maximum Likelihood Estimation (MLE)

A method of estimating the parameters of probability distribution by maximizing a likelihood function, in order to increase the probability of occurring the observed data. We can find MLE by differentiating the above equation with respect to different parameters and setting it to be zero. For example, the derivative with respect to one of the component of parameter alpha i.e. a_j is given by:

$$\frac{\partial l}{\partial \alpha_j} = \sum_{i=0}^{n} \big(y_i - p(x_i; \alpha_0, \alpha)\big) x_{ij}$$

Another name for the logistic function is a sigmoid function and is given by:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

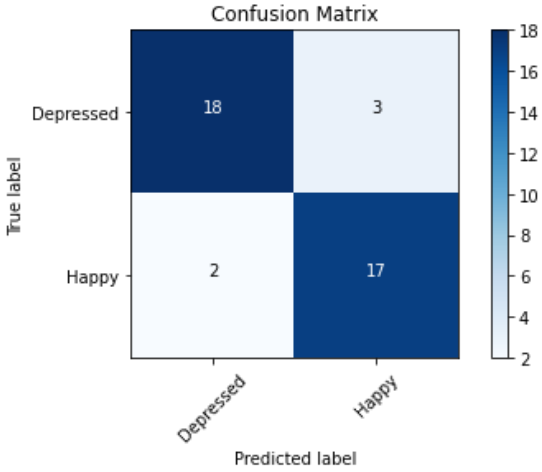
This function assists the logistic regression model to squeeze the values from (-k,k) to (0,1). Logistic regression is majorly used for binary classification tasks; however, it can be used for multiclass classification.

- **For this project we just used the binary Logistic Regression model as it predicts if the person do have depression or not i.e. it gives answers in 0 and 1.**

Chapter 3

3.1 Results & Discussion:

Fig-3.1: Confusion Matrix



The matrix reflects accuracy of the model as **87.5%,** thus the model can be adopted for prediction. The model was able to accurately classify 89% of the not depressed class and 86% of the depressed class.

Fig-3.2: Accuracy and Precision of the model

```
In [148]: from sklearn.metrics import precision_score
          from sklearn.metrics import recall_score
          print("Accuracy:",accuracy_score(y_test, y_prediction))
          print("Precision:",precision_score(y_test, y_prediction))
          print("Recall:",recall_score(y_test, y_prediction))

          Accuracy: 0.875
          Precision: 0.85
          Recall: 0.8947368421052632
```


Fig-3.3: ROC curve

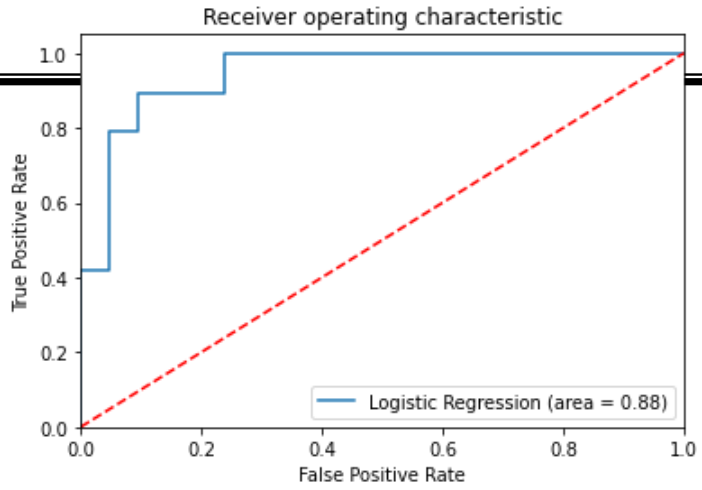
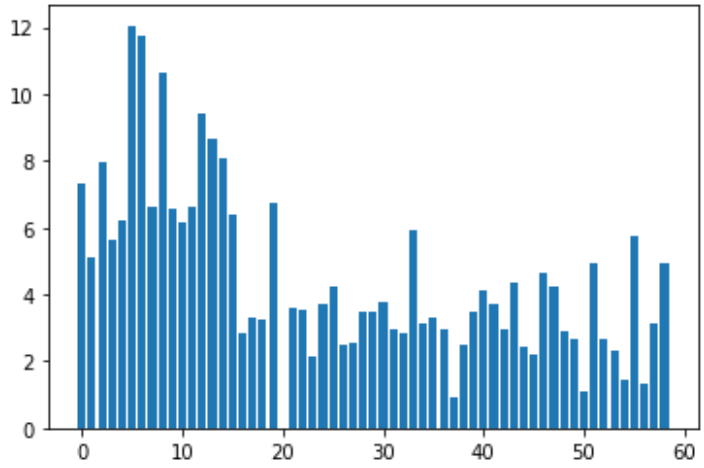


Fig-3.4: features and their significance



Chapter 4

4.1 Conclusion:

It has been possible to predict the mental health of an individual under essentially similar conditions using binary logistic regression equation. The model showed **87.50%** accuracy. The approach adopted for prediction is generic even though model, accuracy of the model, adequacy of the model and predictability of the model will be different for different groups of people subject to essentially similar conditions. Future scope of the study includes considering other models such as Adaptive Boosting(AdaBoost), K-nearest neighbour method(KNN), Gradient Boosting(GB), extreme gradient boosting (XGBoost), Bagging classifier, weighted voting classifier and the likes, comparing the models for the given set of data on performance such as model adequacy parameters viz. AIC, BIC, ROC curve, AUC and the likes, Model accuracy, and predictability of the models using cross validation and selecting the best model.

REFERENCES

[1] Hawryluck L, Gold WL, Robinson S, Pogorski S, Galea S, Styra R. SARS control and psychological effects of quarantine, Toronto, Canada. Emerg Infect Dis. 2004;10(7):1206–12.

[2] Chaturvedi SK. Covid-19, coronavirus and mental health rehabilitation at times of crisis. Journal of Psychosocial Rehabilitation and Mental Health. 2020;7(1):1–2. pmid:32292688.

[3] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–32, 2019.

[4] Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. Artificial Intelligence for mental health and mental illnesses: an overview. Current psychiatry reports, 21(11):1–18, 2019.

[5] Theodoros Iliou, Georgia Konstantopoulou, Mandani Ntekouli, Christina Lymperopoulou, Konstantinos Assimakopoulos, Dimitrios Galiatsatos, and George Anastassopoulos. Iliou machine learning preprocessing method for depression type prediction. Evolving Systems, 10(1):29–39, 2019.

[6] T Nagar. Prediction of mental health problems among children using machine learning techniques.

[7] Adrian BR Shatte, Delyse M Hutchinson, and Samantha J Teague. Machine learning in mental health: a scoping review of methods and applications. Psychological medicine, 49(9):1426–1448, 2019.

[8] M Srividya, S Mohanavalli, and N Bhalaji. Behavioral modeling for mental health using machine learning algorithms. Journal of medical systems, 42(5):1–12, 2018.