Machine Learning for Stress Prediction

by

Sherajus Salehin
19101311
Syeda Tanjima Mahmood
18101165
Muhtasim Fuad Ayon
18101698
Nafiur Rahman
18101366

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
January 2023

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Sherajus Salehin

Sherajus Salehin
19101311

Tanjima

Syeda Tanjima Mahmood
18101165

Muhtasim Fuad

Muhtasim Fuad Ayon
18101698

Nafiur Rahman

Nafiur Rahman
18101366

Approval

The thesis titled "Machine Learning for Stress Prediction" submitted by

1. Sherajus Salehin(19101311)
2. Syeda Tanjima Mahmood(18101165)
3. Muhtasim Fuad Ayon(18101698)
4. Nafiur Rahman(18101366)

Of Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 17, 2023.

Examining Committee:

Supervisor:
(Member)

Annajiat Alim Rasel

Digitally signed by Annajiat Alim Rasel
DN: cn=Annajiat Alim Rasel, o=Brac University, ou=CSE Department, email=annajiat@bracu.ac.bd, c=BD
Date: 2023.01.14 23:04:00 +06'00'

Mr. Annajiat Alim Rasel
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Associate Professor
Department Computer Science and Engineering
Brac University

Abstract

Emotional, psychological, and social well-being are all part of mental health. Stress, social anxiety, depression, and personality disorders are just a few of the elements that build up mental health issues that lead to mental illness. Mental illness is at an all-time high in today's fast-paced world, and it's on the rise. Early detection of mental disorders is critical for preventing mental illness and maintaining a balanced life. Machine Learning (ML) may open up new avenues for recognizing human behavior patterns, as well as detecting irregular mental health symptoms and risk factors. This study gives a systematic view of machine learning approaches to mental health problem prediction. We scan credible resources for research articles and studies relating to machine learning methodologies in predicting mental illness. Machine learning is used in various ways to anticipate mental illness and respond accordingly. Machine learning methods and approaches will aid in the prediction of mental illnesses. To summarize, this thesis attempts to have an impact on the healthcare industry by using machine learning approaches to detect mentally ill patients using large data. We will collect data from the internet through Google form, pre-process the data and use machine learning algorithms to make a model that will predict stress from our selected features. This research work proposes to experiment with various machine learning algorithms (for example scatter matrix plots, decision trees, and logistic regression), compare their performance, and finalize a model to identify the state of mental health status from an organized dataset.

Keywords: Machine Learning, Prediction, Decision tree, Linear Regression, Analysis, mental, stress, anxiety, depression.

Acknowledgement

Firstly, All praise to almighty Allah, for giving us the opportunity to complete our thesis without any major interruption. Secondly, We would like to express our sincere gratitude to our supervisor, Mr. Annajiat Alim Rasel sir, for invaluable guidance, support, and encouragement throughout our thesis. His expertise and patience have been an inspiration to us. We are also grateful to our doctor, Dr. Mohd. Siraj Uddin, for providing us with the necessary medical knowledge and resources to complete our research.

Additionally, we would like to thank all of the data donors who have generously provided us with the information needed to conduct our study. Without their participation, this work would not have been possible.

Lastly, thanks to all our teachers/instructors who have helped us gain valuable knowledge and insights throughout the four years of our under-graduation Life.

We are deeply grateful for all the time, effort, and support that our supervisor, doctor, and data donors have provided us.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mental illness can be defined as a state of mind where the mind is insecure enough to go on with consistent activities. Several predicaments like stress, anxiety, depression and most importantly long-term physical sickness are most likely to fabricate a troubled environment for an individual.

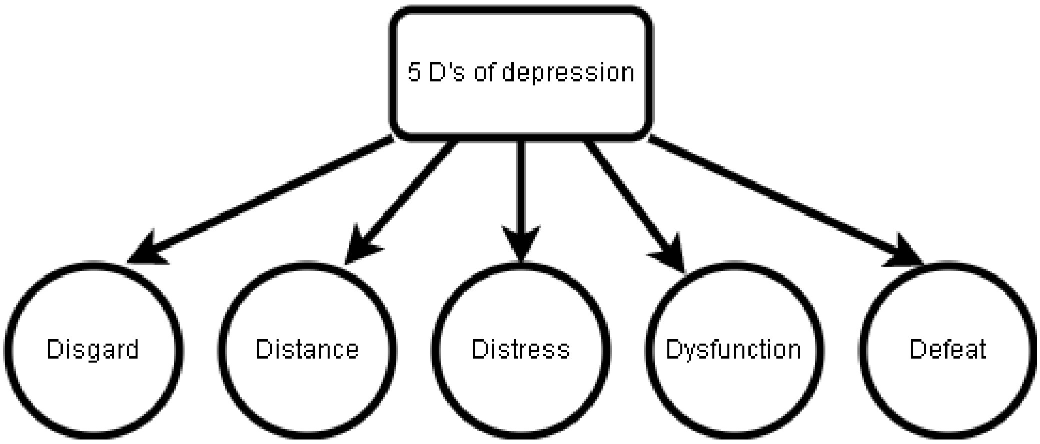


Figure 1.1: 5 D's of Depression

1.1 Stress Types

Firstly, stress can be classified into various types where factors like environment, society etcetera are involved extensively. Some definitions of such stress types can be found in the later sections.

Environment Stressor: Stress that is caused by outside force is most likely to be an environmental stressor. The situation can be described by when an individual is incompetent to reciprocate to both inner and outer state of affairs, which eventually results in stress. For instance, harsh weather conditions, traffic congestion, escalation of criminal offenses, environmental pollution, global pandemic etcetera which contribute greatly to this category.[17]

Physiological Stressor: Each person lives in a different situation compared to others. Everybody has to abide by some social roles, including carrying out a different personality with different people. To stick to so many roles perfectly can bring sudden changes to the mood. As a result, the body and mind are affected severely, which results in physiological stress.[17]

Acute Stress: This is basically a short-lengthed stress, which means the indicators can reveal themselves rapidly, but do not last for a longer period of time. There is also Episodic acute stress that happens at a certain time like teenage period where education makes people stressful without any specific reason. On the other hand, chronic stress can be menacing to health. Generally, this stress may linger for years to come.[17]

Classifications of stress

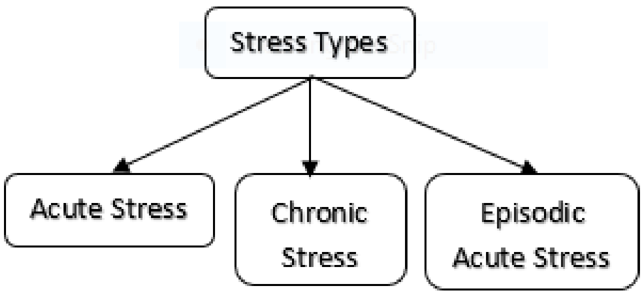


Figure 1.2: Stress classification

1.2 Initial plan

In this section, we will talk about some of the remedies that we can execute in order to diminish the possibility of mental illness being left undiagnosed.

Figure 1.3

In this section, we talk about our plan and the remedies that we can execute in order to diminish the possibility of mental illness being left undiagnosed.

Firstly, the internet is an essential part of everyone's lives nowadays. Daily activities are done through the internet mostly right now. So, it is helpful by a specific margin. Nonetheless, the amount of usage varies from person to person. Excessive utilization can contribute to physical and mental stress. In the year 2021, 4.88 billion people around the planet used the internet, which counts to almost 62 percent of the total population. Among this population, most of the people belong to the young generation to whom the internet is easily available. Furthermore, depression and stress are common issues among young people. The correlation between these numbers is alarming for us. Therefore, to lessen the issue and predict the illness at an earlier stage, we need to work on a balance while using the internet. We will know about this in details in the later parts of our work.

Then comes the survey, which is mostly done through Google form. There will be a specific set of questions that will be asked to the respondents regarding their daily activities, which may assist in revealing a hidden disorder. A survey plays an important role as it helps to understand about what the majority of the people face to be in a stressful situation.

Moreover, Google form authentication setting will be used to limit a person to add his/her data only once. It will help our data to be more balanced.

After that, data preprocessing comes along, which is a process of assembling raw data and forming it fittingly for a machine learning model. It is an important step in machine learning because the qualification of data and the things that can be derived from it, affects the learning ability of our model. There are some steps of completing this preprocessing technique. They are:

1. data quality assessment
2. data cleaning
3. data transformation
4. data reduction

After we have gathered our desired data from the respondents, we will perform the mentioned steps using our model to predict whether they have the illness or not.

A Scatter Matrix Plot can be defined as a grid of scatter plots that are used to envision bivariate relationships between the combinations of variables. There is a chart visible where we can see the aforementioned relation. For our proposed methodology, we will use this plot algorithm to examine the data collected and view them in graphical form, which will help in analyzing and predicting any concealed illness.

Lastly, we will extensively use various machine learning models. A machine learning model can be stated as a file which has been trained to perceive certain types of patterns.Trying multiple algorithms will also help us choose a perfect algorithm for our model. Machine learning is a part of another concept called Artificial Intelligence. It builds a model based on given data as features, also known as training data, which can be used to make predictions or decisions without being explicitly programmed to do so. Therefore, it can be very efficient for our proposed job.

1.3 Research Problem

1.3.1 Lack of Response

Machine learning research is becoming more and more prevalent every day, and the research's limitations are growing as well. According to [8], a training model can be built from a much more detailed and large dataset, but the number of replies is the fundamental constraint. Because most people, even when under stress, strive to avoid answering questions and taking any form of a survey.

1.3.2 The effect of time

According to [8], Between 2017 and 2019, 39,776 cases were collected through online questionnaires using various methodologies. On a separate dataset from DASS21, the same categorization procedures were used.349 people from all over north India filled out Google forms to generate this information. As a result, correctly applying the strategies takes a lengthy time.

1.3.3 Individual Variation and data labeling

According to [5], Interindividual variability is the subject of the approach. It's crucial to note the nature of the person because it's person-specific. As a result, each person must provide some labeled data for data training. Because proper data training necessitates a large amount of data, there is insufficient data. This was one of their most significant limitations. Another drawback was that training data requires at least 15 days per person. It's also necessary for new information or the input one. As a result, it took at least 15 days to adequately train the data. As a result, it is time-consuming. According to [5], For real-world applications, consistent performance is essential. But the previous techniques or methods often lacked the necessary precision. As a result, it makes it difficult to learn more about it.

1.3.4 Over-fit Data

: Several limitations exist in this research work. To begin with, these models use nationally representative survey data. clinical usage of such models might be overfit for specific locations. Secondly, CLHLS data excludes medical use and clinical indicator of these models for the elderly people. After that, due to the limits of the questionnaire format and the amount of CLHLS waves, the LSTM model is unable to use more wave data to forecast risk factors for depression in the elderly, lowering the risk factors' predictive performance. To conclude, the capacity to detect geriatric depression is reliant on medical resources sourced from local healthcare communities. Indications and clinical criteria differ amongst emergency departments and practitioners [14].

1.3.5 Error Correction

: According to [11] this research, the datasets they worked with had a variety of constraints, largely implying the need for “bigger datasets” to make up for missing data, as well as being aware of noise errors while recording data. Due to limitations in the study sample and uncertainty, the established results were also acknowledged as having limited generalizability [11]. Furthermore, issues concerning skewed, missing, or incomplete data were raised in this study [11]. Occasionally, the writers made an error analysis to identify how statistical data was manipulated [11]. A number of recordings clearly detailed difficulties with integrating different, frequently multi-modal data sources, in addition to data processing for desired ML goals and methodologies utilized according to this research work. The lack of a sham condition, a small sample size, and a sparse electrode array are all limitations of research. Despite these methodological flaws, we were able to find results that were both validated and clinically relevant [7].

1.4 Methodology

Machine learning is a way to train a model to do specific tasks / analyze data without human interference. It uses provided statistical data and some preloaded (user selected) algorithms to analyze and make useful patterns for further analysis/ predicting behavior of that dataset. So, we can use machine learning to learn from people’s behavior on the internet (social media, academic/official media, et cetera) and its combination with their mental health status will help the machine to develop a pattern to predict whether a person is mentally ill or not.

Primarily, we will build a dataset containing people’s internet consumption, how they use it, and for how long. Then we will develop a set of questions to learn about their mental health. This processed data will be stored in a database for the machine to analyze.

Then we will have to test and develop algorithms to find relationships between usage and mental health. Doing this thing might require a Correlation heatmap, as it shows how much one variable/relationship affects the other [15].

It will also help relationships to be visualized for human supervision. Preciseness level will be checked by building a Decision tree[9], K-Nearest Neighbor, and the Random Forest algorithm[9].

After that, the algorithm will be able to draw an inferrable pattern for us to test and execute. The whole process will go through supervised learning algorithms only.

1.5 Research Objectives

The goal is to use machine learning in order to create a mental illness prediction system. Typically, we use internet usage statistics as input, as well as user feedback via surveys such as Google Forms. Then it checks the cookie ID and compares it to the records before sending it for further processing via data pre-processing. Then, using a machine learning model, we will build a scatter matrix graphic.Following that, we will create a decision tree based on the results of machine learning approaches.

Thus, we might be able to forecast mental health by using the above procedure. The objectives of this research are

1. Understanding ML (Machine Learning) and how it works.

2. Research the various methods for measuring stress.

3. Collecting data from multiple sources in the form of surveys.

4. Learning about data analysis.

5. To develop and design an effective approach for measuring stress.

6. Experiment with different types of Machine learning (ML) algorithms.

7. To generate recommendations on improving the model.

Chapter 2

Literature Review

2.1 Machine Learning (ML)

Machine learning (ML) is a subfield of artificial intelligence that is described as a machine's ability to emulate intelligent human behavior in a wide sense. The study of machine learning is the study of computer algorithms that can learn and develop on their own through experience and data. It was coined by AI pioneer Arthur Samuel in the 1950s as "the branch of study that enables computers to learn without being explicitly taught." [13]

2.1.1 ML Architecture

Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the four categories in which machine learning is classified [10].

A. Supervised Learning: When faced with new instances, supervised learning happens when an algorithm learns from sample data and associated target responses, which can be numeric values or string labels, such as classes or tags, in order to predict the proper response later [10]. In the training dataset, there are target variables named label and output vector [10]. In the context of supervised learning, there are two types of algorithms [10]. They are regression and classification, respectively [10].

I **Regression**: Linear regression and logistic regression are two types of regression algorithms [10]. Linear regression is a statistical technique for analyzing the relationship between a set of "explanatory" factors and a set of real-valued outcomes [10]. While logistic regression is used to describe the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables, it is also used to describe the relationship between two or more nominal, ordinal, interval, or ratio-level independent variables [10].

II **Classification**: Decision trees, Support Vector Machines (SVM), Naive Bayes, K-Nearest Neighbors (KNN), Random Forest, and Artificial Neural Networks are all classification algorithms [10]. The decision tree is a categorization system that uses a collection of if-then rules to classify items **article1**. SVM is a classifier that uses a statistical learning approach to identify the best hyperplane to categorize the data, whereas Naive Bayes is a classifier that uses an ideal hyperplane to categorize the data [10]. KNN assigns new point values based on feature similarity utilizing various distance functions such as Euclidean distance, Hamming distance, and more. A random forest algorithm is made up of a collection of trees, each of which provides a categorization. For data classification, ANN is based on a human neuron model [10].

B. Unsupervised Learning: When an algorithm learns from basic samples without receiving a response, it is referred to as unsupervised learning. This allows the algorithm to assess specific data patterns [10]. The lack of a supervisor or target variables in the training dataset is referred to as unsupervised learning [10]. This challenge will be solved by the algorithm identifying similarities between the cases [10]. Unsupervised learning is similar to distinguishing a type of chili based on size, color, flavor, and degree of spiciness based on observation [10]. Chili is classified according to how similar it is to other foods [10]. Clustering and dimensionality reduction are the algorithms for unsupervised learning [10].

C. Semi-Supervised Learning: Semi-supervised learning (SSL) is a learning method that uses both classified and unclassified data [10]. It's a combination of supervised and unsupervised learning [10]. SSL aims to address both the benefits and drawbacks of supervised and unsupervised learning [10]. When compared to supervised and unsupervised learning, SLL could offer higher insight because unlabeled data patterns, as well as identified data patterns, boost accuracy [10]. Transductive learning and inductive semi-supervised learning are the two kinds of SSL **article1**.

I **Transductive learning**: Transductive learning is used to predict the exact labels for data that has not been labeled [13].

II **Inductive semi-supervised**: It generates labels for unlabeled data while also producing a classifier [10].

D. Reinforcement Learning: Reinforcement learning is a type of learning that involves taking actions to maximize reward or decrease risk based on observations obtained through interactions with the environment [10]. Observation examples such as errors or false cases are used to teach reinforcement learning. It improves its knowledge of the model by learning from errors [10].

2.2 Related Works

This section will examine previous significant work in the subject of Mental Ill Prediction critically. The diverse methodologies employed for the primary results gained are analyzed, and we illustrate how mental ill prediction utilizing machine learning has its own set of problems, with research work limitations making the embedding of mental ill prediction systems more difficult.

In paper [17], the authors used PSS scale (Perceived Stress Scale) which is then applied to the data which is collected from the university students through primary and secondary sources such as email, Google Drive survey forms. There are various types of questions in PSS scales (Perceived Stress Scale), ranging from emotion to previous month history. The data set was created by PSS (Perceived Stress Scale) to gather an exam that includes a variety of questions, including the entire enthusiastic inquiry. For stress calculation, each choice has its own weight age. Normal, moderate, and high stress are the three levels of stress. For stress calculation, each question has its own weight age. Normal, moderate, and high stress are the three levels of stress. The authors used Perceived Stress Scale (PSS) because it is a tool for assessing stress levels. This scaled display depicts your emotions and sensations over the preceding many months and into the present. Adult ADHD Self-Report Scale is used for asking different types of questions which are related to emotions, feelings, present and past history as it is a stress evaluation test that uses a self-report scale and WEKA is used for the mining of various opinions of people so that it is widely utilized in the opinion mining industry, as well as in other fields.

According to the research [17], the authors primarily used classification algorithms such as Naive Bayes Linear Regression, Multilayer Perceptron, Bayes Net, J48, and random forest, as well as PSS and adult ADHD questionnaires dataset collection, pre-processing, cleaning, feature extraction, and comparison on the basis of their performance parameter, and calculated the accuracy using the performance parameter.

In paper [10], the authors looked at how machine learning may be used to analyze and predict mental health problems, as well as the problem itself and its contributing components. The findings of the paper will be used in other studies to expand on the discussion of mental health issues and their application utilizing computational modeling. According to the research, the most popular data mining technique for resolving challenges involving mental health concerns classification is supervised learning. Support vector machine (SVM) is the most well-known method, followed by decision tree and neural network. These three models have a high accuracy of more than 70%. These three models have high generalization capabilities, thus they don't over-fit.

In paper [6], In software engineers, stress has been highlighted as a major issue. The OSMI (open sourcing Mental Illness) Survey dataset 2017 from the tech industry was used by the authors. The authors use different types of machine learning approaches such as Logistic Regression, KNN Classifier, Decision Trees, Random Forest Classifier, Boosting, Bagging to examine stress patterns in working adults and identify the factors that have a significant impact on stress levels. The authors state that after proper data cleaning and preprocessing, they used a variety of Machine Learning (ML) approaches to train their model. The accuracy of the models was determined and compared. Among the models used, boosting had the highest accuracy. Gender, family history, and the availability of health benefits in the job were found as key characteristics that influence stress using Decision Trees. With these findings, businesses may focus their efforts on reducing stress and providing a more pleasant working environment for their employees.

The research work [12] presents the KNN-classification algorithm used to structure the machine learning model. The model produces a two-way classification of the stress level, determining whether the student is stress-free or stressful, as well as a further classification under stressful pupils, determining whether their stress percentages are too low, medium, or high. Each stressed student receives feedback and a recommended solution from the educational institute based on the range of stress levels and the probabilistic factors of stress. The KNN classification method accurately accomplishes this. This machine learning model has a 94 percent accuracy rate. Not only do the authors determine the criteria and measure stress levels, but they also provide each pupil with the best solution for his or her grief. The learner can apply the remedy and endeavor to keep his or her mental balance.

The research work [8] proposes using data from the online DASS42 tool, eight machine learning algorithms have been applied for predicting psychological issues such as anxiety, depression, and stress. Probabilistic, closest neighbor, neural network, and tree-based algorithms are divided into four categories. The same approaches were used on another dataset, DASS21. The hybrid algorithm had a higher prediction accuracy than single algorithms, but the radial basis function network, a type of neural network, had the highest accuracy. Also included in the dataset are 42 questions from the DASS42 standard form.

According to [9], the data for the study was gathered from publicly available datasets on the internet. For better prediction, the data has been labeled encoded. To obtain labels, the data is subjected to a variety of machine learning approaches. These categorized labels will then be utilized to create a model that can predict an individual's mental health. Our target demographic is the working class, defined as persons over the age of 18. The model will then be implemented into a website, allowing it to anticipate the outcome based on the information provided by the user.

In paper [7], TMS (repetitive transcranial magnetic stimulation) is clinically beneficial for major depressive disorder (MDD) and is being researched for other disorders such as post-traumatic stress disorder (PTSD). The author used a mix of machine learning and electroencephalography (EEG) to see if machine learning analysis of EEG coherence could

1) predict clinical outcomes in people with comorbid MDD and PTSD
2) discern if someone has had TMS treatment.

Before and after TMS, the authors obtained resting-state 8-channel EEG (5Hz to the left dorsolateral prefrontal cortex) and also utilized Lasso regression and Support Vector Machine (SVM) to see if baseline EEG coherence predicted outcome and if it altered after TMS.

In paper [3] focuses on the stress caused by work. To obtain the data, a survey was carried out in several sectors. Psychosocial, environmental, and physical aspects were the focus of the study. Non metric (Support Vector Machine and Neural Network) and metric (Support Vector Machine) techniques are used in the analysis. The precision of the metric techniques is good.

In paper [2], the authors utilized a decision tree technique to analyze the data collected from the two tests and concluded that they were inadequate. Students' stress levels at the beginning of the

term or semester, and at the end of the semester. According to the findings, students' stress levels were lower at the start of the semester and increased towards the end.

According to the description above, most researchers use several types of machine learning algorithms to predict mental illness, such as Logistic Regression, KNN Classifier, Decision Trees, Random Forest Classifier, Boosting, and Bagging.

Chapter 3

Work Plan

The purpose of our model is to generate a proper inference table and keep track of how the data changes. The scatter matrix plot will help us supervise, and the cookie system will help to keep track of data changes. The following figure shows a proper visualization of our work plan.

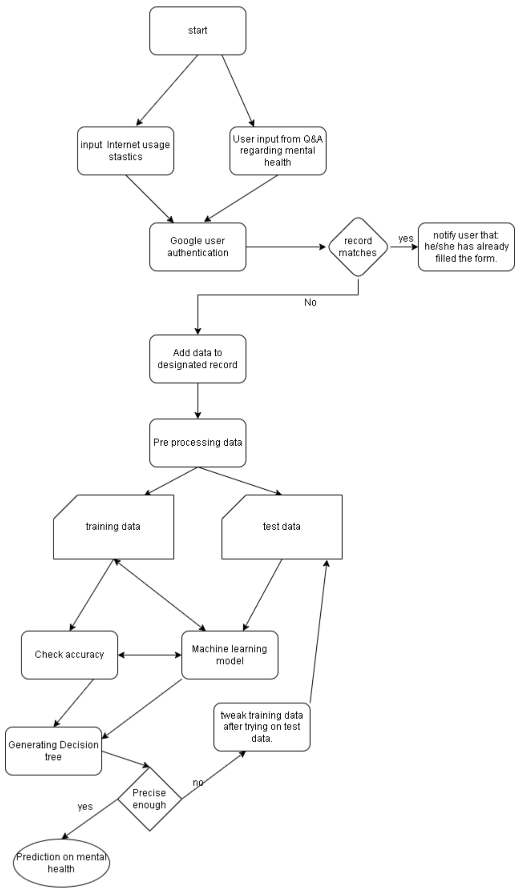


Figure 3.1: Work plan for developing our desired model

The entire work plan can be classified into few stages:

1. **Initialization**: The process starts with inputting necessary data.It has to be processed before directly using it to make a machine learning model.

2. **Tracking data**: To make sure input data goes to their designated slot, we will use Google g sheet generated from Google forms and export their comma separated versions for our use. The form can only be filled once, so that no one can make our data biased in the wrong direction.
3. **Data categorization**: we will categorize data into two parts, training data and test data. We will always test before implementing the actual training data. Another file will be used without any target column (that determines whether someone is stressed or not) for validation purposes.
4. **Processing**: We shall impute/ remove data according to its need. The colab notebook will be reproducible enough for demonstration purposes.
5. **Supervision and related action**: Based on the training model, an accuracy score will be generated. We will tweak the data for best accuracy. The decision tree will go through a random forest method for checking our preciseness level, and then we will change the training data accordingly.
6. **Prediction**: after going through multiple quality checks, the model will be able to predict the mental health status of people based on their internet activity.

Chapter 4

data Input

4.1 Data collection

Finding datasets was not going according to our needs, so we chose to collect data on our own instead.

We are going to use Google form to collect data from people (students) and the user ID based authentication will be handled by google form's authentication system as the form can only be filled once by one user. The form collects two kinds of data. One (age, screen time, social media usage, e-commerce platform usage, online education and professionally identified mental problem) tracks the user's smart device usage habits and another portion (procrastination rate, average pulse and sleep duration) helps to identify whether a person is stressed or not. The second of data will be used to design the target column, which will help us generate a decision tree. Although the second part of data will not take part in pre-processing as they will dominate over other features.The raw data is shown below -

mp	age(numerical- in years)	Gender	Approximated average da	Daily average Number of	Daily average time spent	Daily average time spent	Daily average time spent	Professionally identified m	Procrastination rate	Variation of Pulse per min (you can manually calcul link: Manual to manual He	Daily average sleep duratio
4/2022 15:21:13	23	Male	more than 8 hours [1]	80-120 [2]	2-3 hours [3]	I rarely visit shopping site	more than 4 hours [5]	N/A	very high [6]	80-85 [7]	3-5 hours [8]
24/2022 6:29:09	22	Male	4-5 hours	less than 65	2-3 hours	I rarely visit shopping site	less than an hour	No	medium	75-80	6-8 hours
24/2022 6:44:40	22	Male	4-5 hours	65-80	2-3 hours	0-30 minutes	less than an hour	No	low	70-75	3-5 hours
24/2022 8:49:51	26	Male	4-5 hours	80-120	2-3 hours	I rarely visit shopping site	less than an hour	N/A	low	It mostly stays constant/ c	6-8 hours
24/2022 8:59:17	23	Male	5-8 hours	65-80	less than 2 hours	i don't shop online	less than an hour	n/a	medium	70-75	6-8 hours
24/2022 9:59:37	24	Male	5-8 hours	more than 120	2-3 hours	around an hour	less than an hour		very high	It mostly stays constant/ c	6-8 hours
4/2022 10:16:36	22	Male	5-8 hours	80-120	more than 4 hours	i don't shop online	more than 4 hours	N/A	medium	60-80	6-8 hours
6/2022 11:19:40	24	Male	5-8 hours	less than 65	less than 2 hours	0-30 minutes	more than 4 hours	N/A	low	It mostly stays constant/ c	6-8 hours
4/2022 13:13:48	20	Female	5-8 hours		3-4 hours	I rarely visit shopping site	less than an hour	N/A	high	80-100	8 hours+
4/2022 13:14:05	23	Female	5-8 hours	less than 65	2-3 hours	I rarely visit shopping site	1-3 hours	N/A	medium	80-85	6-8 hours
4/2022 13:14:24	22	Male	5-8 hours	65-80	2-3 hours	i don't shop online	1-3 hours		medium	70-75	3-5 hours
4/2022 13:16:09	20	Male	4-5 hours	less than 65	less than 2 hours	i don't shop online	1-3 hours	N/A	medium	It mostly stays constant/ c	6-8 hours
4/2022 13:22:22	22	Female	5-8 hours	65-80	less than 2 hours	I rarely visit shopping site	3-4 hours	N/A	very high	75-80	3-5 hours
4/2022 13:30:16	22	Male	4-5 hours	less than 65	less than 2 hours	I rarely visit shopping site	1-3 hours	N/A	medium	It mostly stays constant/ c	3-5 hours
4/2022 13:55:07	21	Male	4-5 hours	65-80	less than 2 hours	I rarely visit shopping site	1-3 hours	N/A	medium		8 hours+
4/2022 14:03:28	22	Female	more than 8 hours	65-80	3-4 hours	I rarely visit shopping site	more than 4 hours	N/A	medium	80-100	6-8 hours
4/2022 14:34:23	22	Male	4-5 hours	less than 65	less than 2 hours	I rarely visit shopping site	1-3 hours	N/A	medium	It mostly stays constant/ c	8 hours+
4/2022 15:02:58	22	Female	5-8 hours	less than 65	2-3 hours	I rarely visit shopping site	1-3 hours	N/A	medium	It mostly stays constant/ c	3-5 hours
4/2022 15:47:05	19	Male	more than 8 hours	80-120	less than 2 hours	0-30 minutes	less than an hour	N/A	very high		6-8 hours
4/2022 16:09:29	24	Male	more than 8 hours	less than 65	less than 2 hours	0-30 minutes	1-3 hours	N/A	medium	70-75	6-8 hours
4/2022 19:27:25	22	Male	4-5 hours	less than 65	2-3 hours	i don't shop online	1-3 hours	N/A	medium	60-80	6-8 hours
4/2022 20:10:48	21	Male	5-8 hours	65-80	3-4 hours	I rarely visit shopping site	1-3 hours	N/A	low	It mostly stays constant/ c	6-8 hours
4/2022 21:05:55	22	Male	4-5 hours	more than 120	less than 2 hours	I rarely visit shopping site	1-3 hours	N/A	medium	80-100	8 hours+
4/2022 21:06:10	22	Male	5-8 hours	80-120	2-3 hours	I rarely visit shopping site	less than an hour	Mild Personality Disorder	high	80-100	6-8 hours
4/2022 21:07:26	20	Male	more than 8 hours	65-80	2-3 hours	I rarely visit shopping site	1-3 hours	N/A	very high	80-100	8 hours+
4/2022 21:11:05	21	Male	4-5 hours	less than 65	2-3 hours	I rarely visit shopping site	1-3 hours	NA	medium	75-80	6-8 hours

Figure 4.1: Raw data

4.2 Data Annotation

Since we collected and categorized our own data based on some internet sources (credible), We decided to show and validate our data from a professional doctor.

In our case, the final model consists of a class identifier that uses Heart rate variability to predict stress. Heart rate variability(HRV) is the variability of time between two pulses. Here, low variability means higher stress level and high variability means lower stress level.

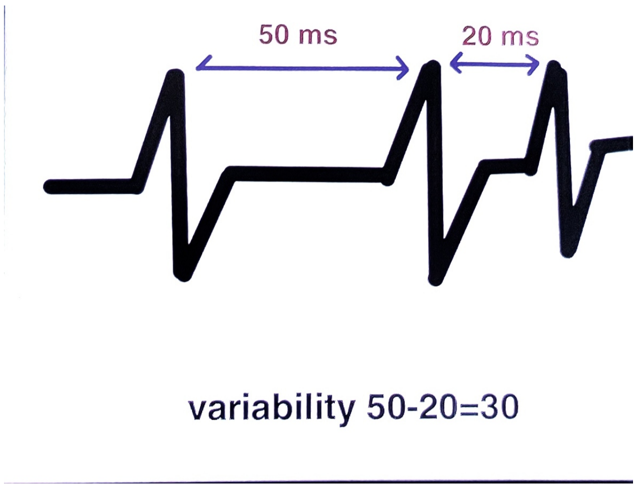
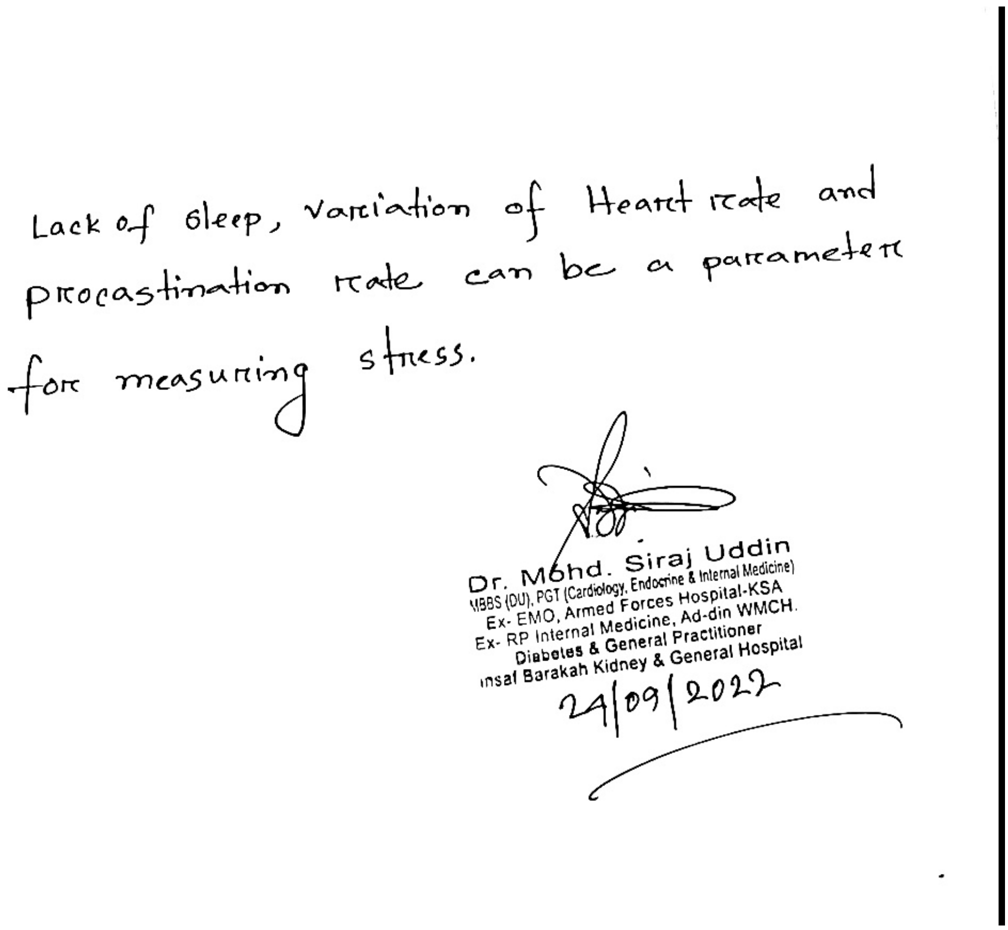


Figure 4.2: Example HRV calculation

Since, our data donors mostly rely on manual / smart band based data to measure HRV, we asked for few measurements of heart rate (pulse) within negligible interval for three minutes and return us the average fluctuation with a higher and lower band. Instantaneous Heart rate measurement can show the interval between beats. for example, If the pulse is 80, then the interval between beats will be $\frac{80}{60} = 1.3 seconds\ 1300 milliseconds$. Mathematically,
$Let, twopulses'x'and'y'$
$So, interval1equals\frac{x}{60}$
$andinterval2equals\frac{y}{60}$
$ThustheHRVoffirsttwopulses = |\frac{x}{60} - \frac{y}{60}|$ Here we can see that HRV changes according to the change in instantaneous HR. Thus, a continuous stable HR measurement means low HRV and instability means high HRV. So, we have showed our measurement technique and the collected dataset to a cardiology specialist and asked for feedback in the form of a statement, so that we can annotate our data reliably.



Lack of sleep, variation of Heart rate and procastination rate can be a parameter for measuring stress.

Dr. Mohd. Siraj Uddin
MBBS (DU), PGT (Cardiology, Endocrine & Internal Medicine)
Ex- EMO, Armed Forces Hospital-KSA
Ex- RP Internal Medicine, Ad-din WMCH.
Diabetes & General Practitioner
Insaf Barakah Kidney & General Hospital
24/09/2022

Figure 4.3: Data annotation

Chapter 5

Pre-Processing

Before pre-processing via python libraries, we need to manually assign the target values via a spreadsheet. Target values will be determined by using user's procrastination rate, heart rate variation and sleep duration. Stressed people tend to procrastinate more than average people[1]. The Heart rate can also predict stress level by its variation range. The more variation, The more stress[1]. And any sleep duration rather than the 6-8 hour range is a sign of stress, but normally stressed people tend to sleep less than 5 hours. Now a stressed person might check all the upper mentioned factors, but we shall have 2 out of 3 irregularities as value 1 for target column and 0 if the data says otherwise. However, Heart rate variation(HRV) is an independent identifier for stress detection, so in case of less accuracy, we shall use HRV for building target columns only. Initial pre-processing starts in google spreadsheet(G-sheet), that encodes text based data to numerical data. Raw data and their encoded numeric format are given below.

Data

The whole pre-processing part will go through some steps. 1.Level encoding: In this stage, easily comprehensible user inputs (attribute of features) will be changed to numerical Levels. Changed levels are shown below.

Due to the usage of long column names, I shall replace them with their shorter code names. Here is the list of shortened column names:

1. Gender→ G
2. Approximated average daily usage of smartphone and personal computer→Screen Daily average Number of Unlocks of smartphone (if available)→ Unlock
3. Daily average time spent in social media → sm
4. Daily average time spent in Shopping sites → sh
5. Daily average time spent in online educational sites → edu
6. Professionally identified mental sisorder → md

The encoded versions of all features will contain capital 'E' in their designated columns.

Table 5.1

G	EG	Screen	EScreen	Unlock	EUnlocks	sm	Esm	sh	Esh	edu	Eedu	md	Emd
Female	0	0-3 Hours	1	Less Than 65	1	I don't use social media	0	I don't shop online	0	Less than an hour	1		
Male	1	4-5 Hours	2	65-80	2	Less Than 2 Hours	1	I rarely visit shopping sites	1	1-3 Hours	2		
Other	2	5-8 Hours	3	80-120	3	2-3 Hours	2	0-30 Minutes	2	3-4	3	To Be Determined Later	To Be Determined Later
		More Than 8 Hours	4	More Than 120	4	3-4 Hours	3	Around an Hour	3	More Than 4 Hours	4		
						More Than 4 Hours	4	More Than an Hour	4				

Encoded Data Preview:

age(numerical- in years)	Gender	Approximated average dai	Daily average Number of U	Daily average time spent i	Daily average time spent i	Daily average time spent i	Professionally identified m	Procrastination rate	Variation of heart rate.	Daily average sleep durati
22	1	2	2	2	2	1	0	0	0	2
26	1	2	3	2	1	1	0	0	0	0
23	1	3	2	1	0	1	0	1	0	0
24	1	3	4	2	3	1		3	0	0
22	1	3	3	4	0	4	0	1	1	0
24	1	3	1	1	2	4	N/A	0	0	0
20	0	3		3	1	1	0	2	1	1
23	0	3	1	2	1	2	0	1	0	0
22	1	3	2	2	0	2		1	0	2
20	1	2	1	1	0	2	0	1	0	0
22	0	3	2	1	1	3	0	3	0	2
22	1	2	1	1	1	2	0	1	0	2
21	1	2	2	1	1	2	0	1		1
22	0	4	2	3	1	4	0	1	1	0
22	1	2	1	1	1	2	0	1	0	1
22	0	3	1	2	1	2	0	1	0	2
19	1	4	3	1	2	1	0	3		0
24	1	4	1	1	2	2	0	1	0	0
22	1	2	1	2	0	2	0	1	1	0
21	1	3	2	3	1	2	0	0	0	0
22	1	2	4	1	1	2	0	1	1	1
22	1	3	3	2	1	1	Mild Personality Disorder	2	1	0
20	1	4	2	2	1	2	0	3	1	1
21	1	2	1	2	1	2	0	1	0	0

Figure 5.1: Encoded data preview

5.1 Missing data imputation

There are some scopes in our data collection form, where people can skip answering questions. So, null values are expected. Thus, the next step needs to impute missing values. Features that can contain null values are: Gender, Daily average Number of Unlocks of smartphone, Daily average time spent in online educational sites, Professionally identified mental/neurological disorder (write 'N/A' if you don't have any), Variation of Pulse per minute/ heart rate, Daily average sleep duration. Firstly, we will use Level Encoder to impute 'Gender', then we might need to scale using min-max scaler for imputing Unlock count and other null values will be handled by Level encoder. But as it turns out, only 'Number of Unlocks' and 'Professionally identified mental disorder' have null values.

```
data.isnull().sum()

age(numerical- in years)                              0
Gender                                                0
average daily smartphone and personal computer usage  0
Number of Unlocks                                     2
time spent in social media                            0
time spent in Shopping sites                          0
time spent in online education                        0
Professionally identified mental sisorder             2
target                                                0
dtype: int64
```

Both have been imputed by the median strategy of Sklearn's SimpleImputer library.

5.2 Scaling

Next stage uses a Standard Scaler to scale data and eliminate any dominating feature that might make the decision biased.

5.3 Implementation and Result

We have tried multiple methods(Random forest, k-Nearest Neighbor, Gaussian Naive Bayes) to make our model a multi-class classifier that will predict whether a person is stressed or not along with stress level. But whatever algorithm we use, multiclass accuracy never exceeds over 57%. Changing the test-train split ratio does not help and the accuracy does not move up when we give it more samples. Since accuracy is important here, we dropped some features that generate confusion(Based on feature importance(a module from Random forest classifier library))

Besides, less important features, we have dropped age as our collected data is biased towards 22 year.

However, our Heart Rate Variation column, identifier of Stress (target) has good variety and should work great for a binary-classification model. Thus, we decided to try a Logistic regression model, and it gave us an accuracy of 80%. To validate our accuracy, we exported a separated data without any target (class) column and

	feature	importance
3	Daily average time spent in social media	0.227254
1	Approximated average daily usage of smartphone...	0.179887
4	Daily average time spent in Shopping sites	0.168297
2	Daily average Number of Unlocks of smartphone(...	0.156964
5	Daily average time spent in online educational...	0.149359
0	Gender	0.072886
6	Professionally identified mental sisorder	0.045353

Figure 5.2: Feature Importance

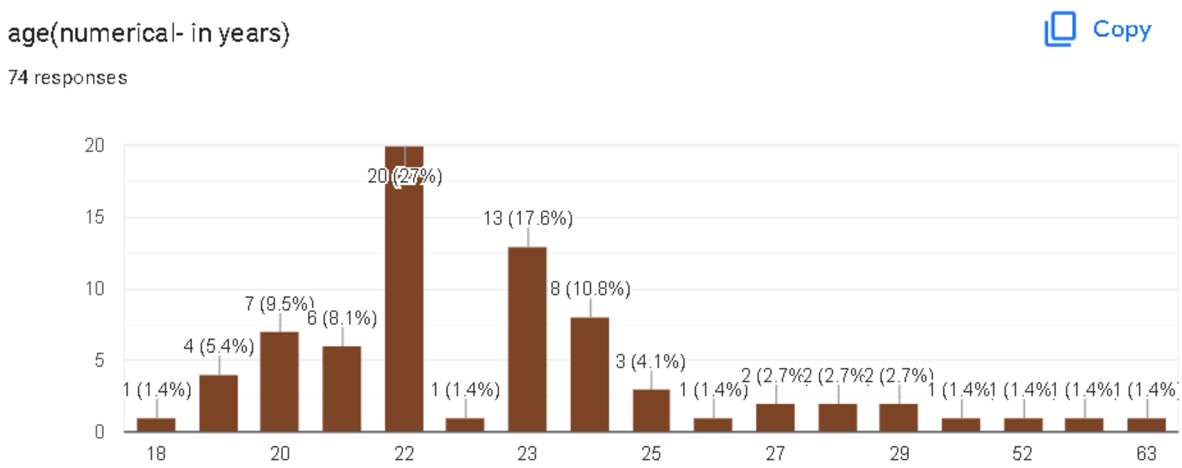


Figure 5.3: Quantity of Responders of Same Age

used it in our model to predict binary values (0:normal, 1:stressed) that identify whether a person is stressed or not. During our test, we have also found few false negatives(0 for stressed people) but 80% accuracy feels good enough to implement a primary stress identifier. We still decided to build another Decision tree classifier for our binary-classification problem, for an easy interpretation of how the identification process works.

Decision Tree (figure 5.3) preview (click the figure title for better resolution picture):

According to this tree, Smartphone unlock count is the purest node(GINI=0.346) and works as the root node. Data below the unlock count threshold(0.407) has more non-stressed class leafs than the other side. So, it can be said that unnecessarily unlocking a phone can be an identifier of stress. Besides, most leaf nodes come from a decision node of either 'unlock count' or 'screen on time'. Both result in class1(stress) for higher value and class0(normal) for lower value. Some decision branches decide based on 'time spent on social media/ online shopping sites' and very few leaf nodes rely on whether a person has any 'professionally identified mental problems' or not. However, 'time spent on educational platforms' reacts differently against other features. Higher values in this feature yield class0(normal) and class 1 otherwise. Figuratively, from this tree, we can say that- "getting work done helps people to stay normal."

5.4 Result Analysis

We have tried various types of learning methods and also tested data with fewer features on them to reduce confusion. The following table represents a timeline of our experiment that yields implemented algorithm names, their accuracy and classification type.

Table 5.2: Accuracy timeline

Algorithm Name	best Accuracy	Classification type
Random Forest	55%	multi-class
Random Forest(fewer features)	19%	multi-class
Gaussian Naive Bayes	25%	multi-class
Random Forest(added feature-sleep duration)	57%	multi-class
Logistic Regression(Selected one)	80%	Binary-classification
Decision Tree	70%	Binary-classification
Logistic Regression(fewer features)	87.5%	Binary-classification

Thus, we can say that binary classification (Logistic Regression) is the best algorithm for our desired model. But in future, if we manage to get more qualitative and quantitative data, we shall try again to predict stress levels using multi-class classification types.

The following links contain the address to our Notebook and Data-set. The notebook

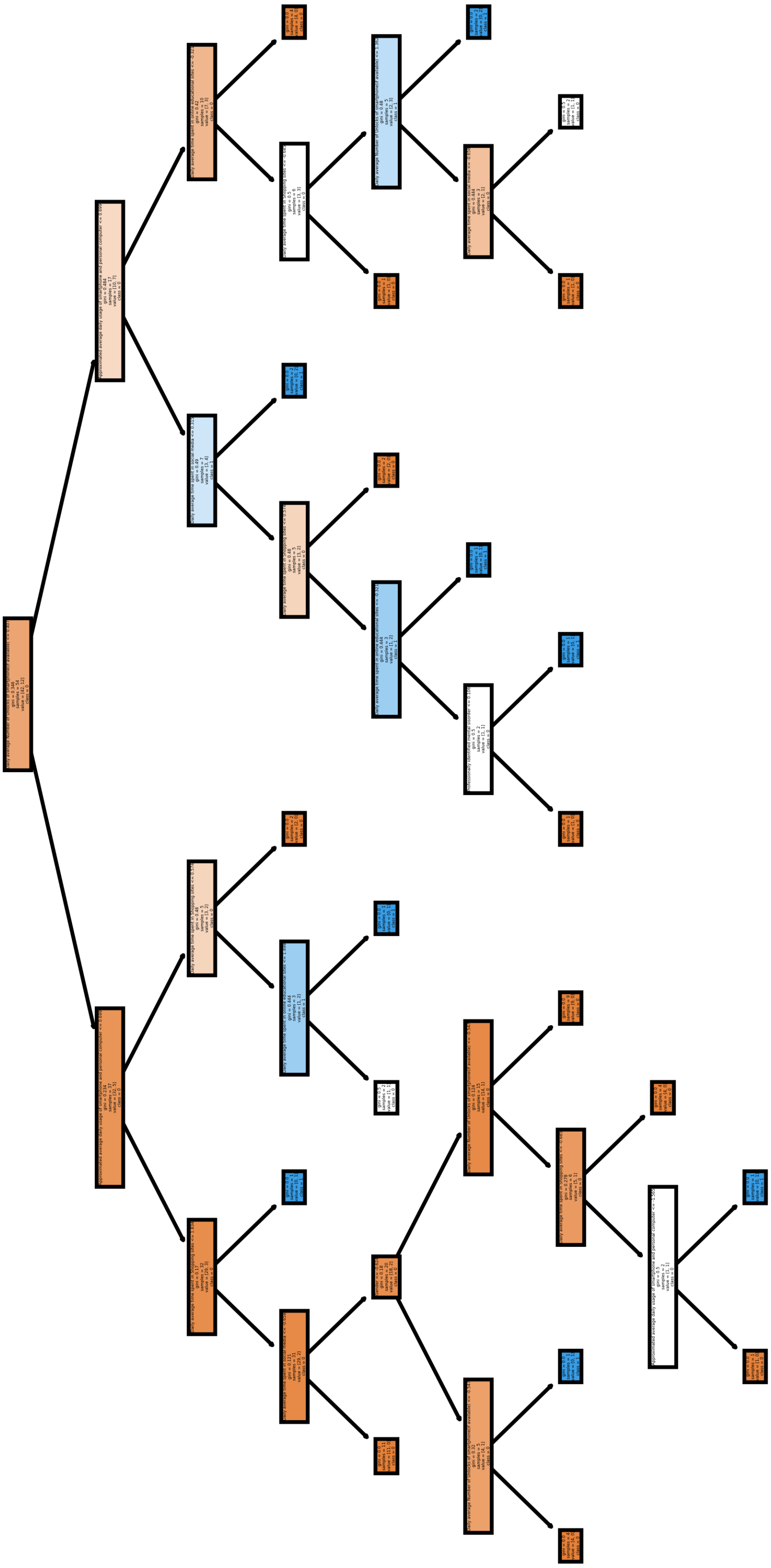


Figure 5.4: Decision tree

is well organized and comprehensive enough. A “table of contents” is also added there for easy navigation.

Colab File and data Link:

1. Colab Notebook link: https://colab.research.google.com/drive/1nYDXjcDTs20Q_S7xL3G55QUpjHc7fVn2?usp=sharing
2. Link to The encoded Dataset: https://docs.google.com/spreadsheets/d/1Tjit8pv4kcrOGh9Jtny edit?usp=sharing

Chapter 6

Conclusion

Although research on the potential mental health advantages of machine learning has been going on for some time, fascination regarding the topic (specially machine learning) is on a high lately. Moreover, we discussed current trends and challenges in this field of study. When it comes to establishing machine learning systems for mental health therapy, our work has been associated with both established methods and novel pathways for achieving this goal. It is mostly considered as a tool of analyzing data which, in turn, automates analytical model building. As mentioned previously, it connects to the branch of Artificial Intelligence which is based on the notion of systems learning from data, identifying certain patterns and reaching a decision with less human interaction. Statistically, around 300 million people worldwide are depressed, according to the World Health Organization [4]. According to our research, it is incredibly difficult to gather such large-scale, high-quality data for study design. Also, depression will be the leading cause of global illness burden by the year 2030 [16]. So, in order to better understand the particular problems and demands of MHPS and those who have had mental health issues, we have pushed researchers to step up their work in this area. In order to ensure that machine learning can genuinely profit from breakthrough data technologies, it is essential to conduct deeper and more creative explorations of the design space. Diagnostic and treatment planning are two obvious applications of machine learning in mental health treatments. In order to make better decisions about data uses and to increase their confidence and acceptance of data applications that come as a result of data access, it is necessary to help individuals evaluate the potential benefits of data sharing and the extent to which potential risks are mitigated or outweighed by potential benefits (for example, the effectiveness of interventions). Machine learning developments in mental health should be presented cautiously to prevent premature claims about its usefulness and real-world impact because of their subject's infancy. The difficulty and complexity necessary to produce long-lasting and dependable machine learning outcomes emphasizes the significance of this. Because most models have not been tested in clinical settings, there is a dearth of evidence on their effectiveness in improving mental health outcomes or services. Even though machine learning models have frequently been touted to outperform conventional research and clinical procedures, we feel like it is vital to regard these approaches as complementing rather than competing. The use of ML treatments in the future to suit the needs of both persons undergoing mental health care and mental health professionals will require more research. In the steps of data processing, interpretation, and visual representation, it is important to prevent excessive abstraction or translation from the individual and their unique circumstances when using machine learning algorithms that capture a wide variety of human needs and experiences. To be precise, we have utilized many machine learning algorithms to implement the method of predicting mental disease. We have discussed initializing algorithms like Decision Tree, WEKA, Scatter Matrix Plot and many more. The testing part will be conducted through surveys and google forms mostly, though cookie ID may also be needed. We will use Data preprocessing to train our collected data and then implement the process in a webpage after the designing is done. After that, the users will avail the advantage of filling out surveys and forms. Therefore, to conclude, we argued that in order to help the field achieve its many lofty objectives in terms of machine learning in mental health, it is necessary to continue conducting basic, multidisciplinary research in close collaboration with health partners, developing and testing new machine learning interventions, and evaluating their effectiveness in real-world settings. This discussion focuses on the problems of building new machine learning-enabled solutions that are understandable and (clinically) useful to their intended audience. Humans, health care, and society as a whole are impacted by the usage of machine learning systems in a variety of ways, and research and development efforts must account for these effects.

6.1 Future Work

6.1.1 Collaboration

We have observed that a lot of 'smart fitness band' brands have already implemented the stress identification functionality using HRV **r23**.Besides, Google android manufacturers already include the app-" Digital Wellbeing" in their software, that keeps user's screen time data including apps and their usage time **r24**. So, if we ever get the chance to acquire data from wearables and digital wellbeing and merge them together, a more precise model will be generated that will predict stress more accurately without any smart-bands. Currently, we are searching for an effective way to communicate with brands that produce both android and wearables.

6.1.2 Better way to collect HRV

The HRV (Heart Rate Variation) calculation was unclear to lots of data donors. Making it easier could have accelerated our work. To make an upgrade over our data collection method, we have designed a circuit consisting:

1. Development board : ESP 8266 / Arduino UNO R3
2. Sensor : Pulse sensor
3. Display module: I2c
4. Wires: male jumper cables

This circuit will be able to read the heart rate of our data donor and do the calculations for him/ her, making our data collection easier.

https://www.overleaf.com/project/63c4467661d7536453499406

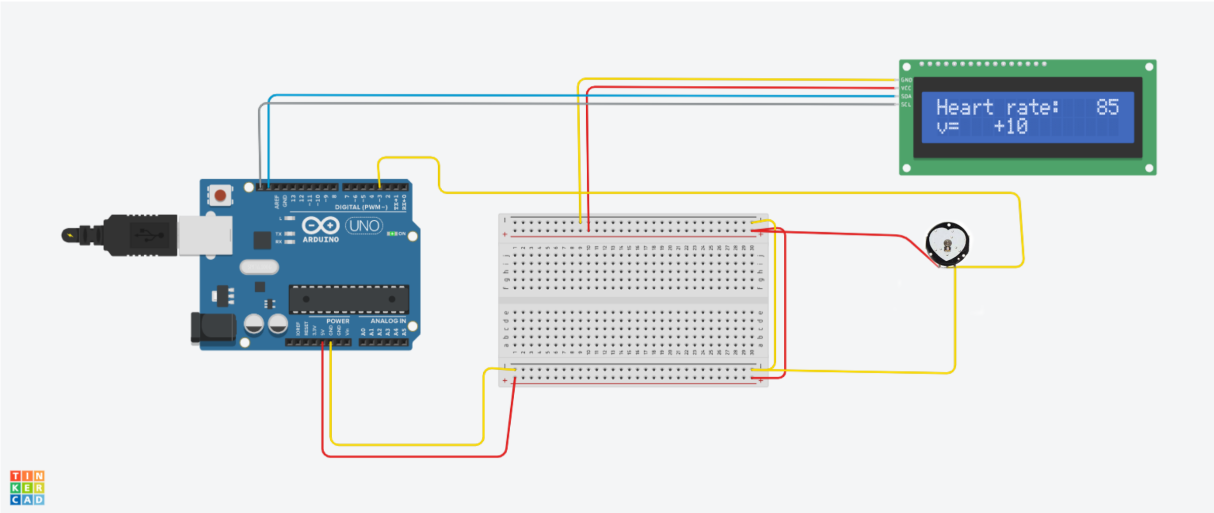


Figure 6.1: Proposed circuit

6.1.3 Image processing

Our future plan consists of taking still photos of human face, classify it between stressed people photos and Non-stressed people photos. Then the classified data will go through some image processing methods (preferably Google v5 network) to make a model that will be able to tell whether a person is stressed or not.

6.1.4 Combining different stress prediction methods

The reason we are trying to measure stress among smart device owners is because it is an easy method to implement a way to measure stress in people. Stress makes a people less productive and thus pushes a person towards failure and depression. If we could monitor stress, handling it would have been a possibility. The shown way is just a skeleton, we can implement similar method for identifying stress in various other ways as well. An education institution might be able to measure stress by monitoring student's attendance, exam appearance, score and many more. The same goes for other offices where efficiency is important. We hope to combine such ideas and implement it.

Bibliography

[1] S. Hasija, 1993. [Online]. Available: https://books.google.com.bd/books?id=nI0xLKVOOYoC.

[2] S. Norizam, "Determination and classification of human stress index using non-parametric analysis of eeg signals," 2015.

[3] Q. Xu, T. L. Nwe, and C. Guan, "Cluster-based analysis for personalized stress evaluation using physiological signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 275–281, 2015. DOI: 10.1109/JBHI.2014.2311044.

[4] Jul. 2017. [Online]. Available: https://archive.dhakatribune.com/feature/2017/07/24/depression-still-considered-weakness-2.

[5] N. Jaques, O. (Rudovic, S. Taylor, A. Sano, and R. Picard, "Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation," in *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, N. Lawrence and M. Reid, Eds., ser. Proceedings of Machine Learning Research, vol. 66, PMLR, 20 Aug 2017, pp. 17–33. [Online]. Available: https://proceedings.mlr.press/v66/jaques17a.html.

[6] U. S. Reddy, A. V. Thota, and A. Dharun, "Machine learning techniques for stress prediction in working employees," in *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2018, pp. 1–4. DOI: 10.1109/ICCIC.2018.8782395.

[7] A. Zandvakili, N. Philip, S. Jones, A. Tyrka, B. Greenberg, and L. Carpenter, "Use of machine learning in predicting clinical response to transcranial magnetic stimulation in comorbid posttraumatic stress disorder and major depression: A resting state electroencephalography study," *Journal of Affective Disorders*, vol. 252, pp. 47–54, Jun. 2019. DOI: 10.1016/j.jad.2019.03.077.

[8] P. Kumar, S. Garg, and A. Garg, "Assessment of anxiety, depression and stress using machine learning models," *Procedia Computer Science*, vol. 171, pp. 1989–1998, Jan. 2020. DOI: 10.1016/j.procs.2020.04.213.

[9] V. Laijawala, A. Aachaliya, H. Jatta, and V. Pinjarkar, "Mental health prediction using data mining : A systematic review," *MatSciRN: Process & Device Modeling (Topic)*, 2020.

[10] N. Shafiee and S. Mutalib, "Prediction of mental health problems among higher education student using machine learning," *International Journal of Education and Management Engineering*, vol. 10, pp. 1–9, Dec. 2020. DOI: 10.5815/ijeme.2020.06.01.

[11] A. Thieme, D. Belgrave, and G. Doherty, "Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems," *ACM Trans. Comput.-Hum. Interact.*,

vol. 27, no. 5, Aug. 2020, ISSN: 1073-0516. DOI: 10.1145/3398069. [Online]. Available: https://doi.org/10.1145/3398069.
[12] P. Manjunath, T. S, P. Shreya, V. Ashok, and D. S. Sultana, "Predictive analysis of student stress level using machine learning," *International Journal of Engineering Research Technology*, vol. 9, no. 12, Jul. 2021. DOI: 10.17577/IJERTCONV9IS12019. [Online]. Available: https://www.ijert.org/predictive-analysis-of-student-stress-level-using-machine-learning.
[13] M. Sloan, *Machine learning, explained | mit sloan*, Apr. 2021. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.
[14] D. Su, X. Zhang, K. He, and Y. Chen, "Use of machine learning approach to predict depression in the elderly in china: A longitudinal study," *Journal of Affective Disorders*, vol. 282, pp. 289–298, 2021, ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2020.12.160. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016503272033250X.
[15] 2022. [Online]. Available: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_scatter_matrix_plot.htm.
[16] [Online]. Available: https://www.wfmh.org/2012DOCS/WMHDay%202012%20SMALL%20FILE%20FINAL.pdf.
[17] D. Sharma, N. Kapoor, D. Sandeep, and S. Kang, *STRESS PREDICTION OF STUDENTS USING MACHINE LEARNING*. [Online]. Available: http://www.tjprc.org/publishpapers/2-67-1597130601-534IJMPERDJUN2020534.pdf.