Post-COVID19 Psychological Distress Detection Using Machine Learning Techniques

Junior Design Course Project (CSE299) Report

Instructor
Dr. Shahnewaz Siddique
Assistant Professor
Dept. of Electrical and Computer Engineering
North South University

Submitted by
Shafayet Rajit 1921325042
Mezbah Uddin Saad 1921703042

Section 4

10th September, 2022

Abstract

COVID-19 has affected people's lives on many dimensions. Human beings could not interact with each other normally for a long time. This, along with various other issues, had a severe impact on people's mental health. In this project, we aim to create a machine learning model which will predict whether one is suffering from anxiety or other mental health problems based on some of the answers they provide to certain questions. This project also focuses on finding out how much COVID-19 has affected our mental health in the long run.

There are two parts of the data collected from numerous respondents, one of that is information regarding personal lifestyle and another is a popular anxiety assessment questionnaire, GAD-7. Using this information, we have trained a machine learning model. This model will take into account the various aspects of personal lifestyle during COVID-19. After that, it will also analyze the GAD-7 score for each respondent. Finally, it will try to predict anxiety levels in individuals using only information on personal lifestyle.

We have used the collected data to train machine learning models utilizing multiple algorithms. Among them, the decision tree algorithm has produced the highest accuracy with a score of 54.7%. This project targets to analyze mental health using highly efficient machine learning algorithms in order to produce rapid and correct detection of psychological distress.

Contents

1 Introduction

COVID19 has affected people's lives on many dimensions. Human beings could not interact with each other normally for a long time. This, along with various other issues, had a severe impact on people's mental health.

In this project, we aim to create a machine learning model which will predict whether one is suffering from depression or other mental health problems based on some of the answers they provide to certain questions. These questions will be related to their lifestyle during the COVID19 pandemic. Observing their lifestyle during the pandemic, the model will show the level of anxiety or other psychological distress one may be going through.

This project is titled "Post-COVID19" as in our country, the number of COVID-19 patients is gradually lessening right now. Even though there are sudden increases, the overall situation is much more bearable than before. So, we will ask the participants questions regarding their situation during COVID-19 and use that information to predict their mental health.

2 Problem Statement

To build a machine learning model that will be able to detect the level of psychological distress one may experience due to their lifestyle in COVID-19.

3 Background

Professionals all over the world have worked to successfully detect the wellness of mental health. There was a spike in research on this field during the COVID19 pandemic as the impact became evident. Machine learning techniques have led to rapid and accurate results in these researches. After collecting the necessary data from willing participants, the researchers trained a machine learning model using suitable algorithms. Some of the popular algorithms are KNN, SVM, Logistic Regression, Random Forrest, Decision Tree, etc.

Prout, Tracy A., et al. (2020) collected data from 2,787 people from different backgrounds and implemented random forrest algorithm to train the model. They have taken into account the demographics, history of adverse childhood experiences, current coping strategies, and current psychological distress. Their findings indicate that younger participants, women, and non-binary individuals reported higher prevalence of symptoms across all measures of distress. [5]

Sau et al. (2017) manually collected data from the Medical College and Hospital of Kolkata, West Bengal on 630 elderly individuals, 520 of whom were in special care. After applying different classification methods they produced a model with the best accuracy rate of 91% and 89% among the two data sets of 110 and 520 people, respectively. [6]

4 Solution Process

4.1 Data Collection

In order to collect data, we prepared a set of questionnaires related to lifestyle during the COVID19 pandemic. There were two parts of the questionnaire, one is about personal lifestyle and another is the GAD-7 questionnaire.

The questions in the personal lifestyle section requested information regarding demography and lifestyle during COVID-19. The detailed questionnaire is mentioned in the appendix section.

The second part of the questionnaire was a set of questions titled GAD-7 questions. This assessment was chosen as it has displayed good agreement between self-report and interviewer-administered assessments meaning this method can be used with confidence in both scenarios.

The original validation study conducted using GAD-7 on adult patients in primary care clinics in USA reported a Cronbach's α score of 0.92 [7]. Other clinical and non-clinical studies conducted in Korea, Portugal, the United States, Iran, Germany, and, Peru have similarly found excellent Cronbach's α coefficient which shows good internal consistency of the GAD-7 scale across different populations [8, 2, 1, 4, 3]. Specifically, surveys conducted on university students in Korea and college students in Portugal found Cronbach's α coefficient to be 0.91 and 0.88 respectively, revealing excellent internal consistency [40, 41]. In Bangladesh, Faisal et al. (2021) found good internal consistency of GAD-7 (Cronbach's $\alpha = 0.87$) in a study on university students [14].

There are seven questions in the GAD-7 questionnaire and with each question, there is a level of severity from 0 to 3. The participant must choose the number applicable to their scenario. Then, after combining all the scores, the final score is used to determine the level of anxiety. The questions and score levels are presented in Figure 1 and Figure 2 respectively.

Over the last two weeks, how often have you been bothered by the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Feeling nervous, anxious, or on edge	0	1	2	3
2. Not being able to stop or control worrying	0	1	2	3
3. Worrying too much about different things	0	1	2	3
4. Trouble relaxing	0	1	2	3
5. Being so restless that it is hard to sit still	0	1	2	3
6. Becoming easily annoyed or irritable	0	1	2	3
7. Feeling afraid, as if something awful might happen	0	1	2	3

Figure 1: GAD-7 Questions

Score	Anxiety Level
0 - 4	Minimal
5 -9	Mild
10 - 14	Moderate
15 - 21	Severe

Figure 2: GAD-7 Results

We have collected the data using an online form, named Google forms, as it is easier to reach a large number of people through this medium. This data collection process was completely anonymous. Any information that may identify an individual has not been requested in this data collection process.

4.2 Data Processing

We have collected 230 responses in total over the course of one month. We aimed to collect categorical data so that it is easier to process. In order to ease the management of this data, we have assigned shortened labels to each one of them. Additionally, we have simplified the categorical data by assigning indicator values to them. After collecting the data, we used visualization libraries, such as Seaborn, to analyze the data and get an overview of the dataset.

In Figure 4, we can see that we have managed to get a balanced representation of both males and females. However, it is evident observing Figure 3 that most of the data came from people who are between the ages of 18 and 24. Similarly, if we look at Figures 5 and 6, we can draw conclusions that most of the respondents are unmarried students. Thus, in terms of demographic information, we can conclude that there is a certain amount of bias in our dataset.

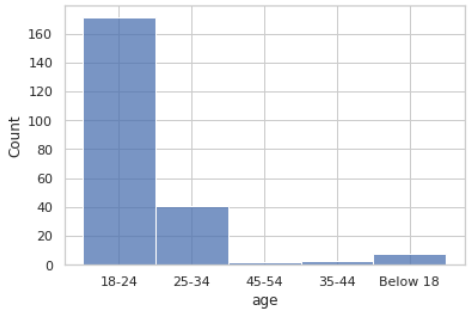

Figure 3: Age

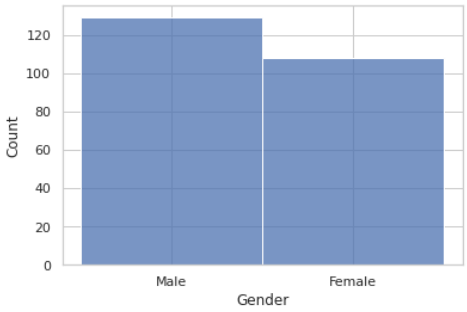

Figure 4: Gender

Now, Figure 7 represents the family monthly income of respondents. We have assigned short terms for better visualization. The scale is stated below:

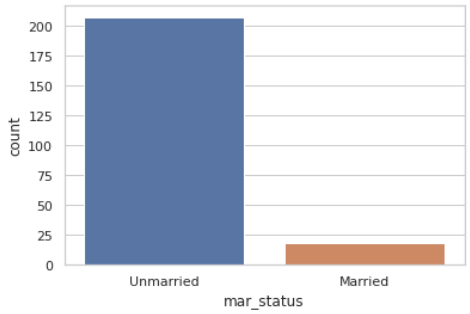

Figure 5: Marital Status

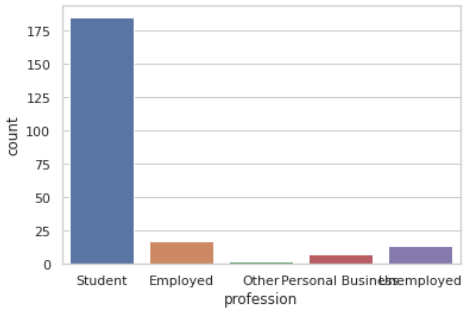

Figure 6: Profession

1. 'More than 1,50,00 BDT':'SRI',
2. '1,20,000 - 1,49,999 BDT':'RI',
3. '1,00,000 - 1,19,999 BDT':'HC',
4. '70,000 - 99,999 BDT':'HMC',
5. '50,000 - 69,999 BDT':'MC',
6. '30,000 - 49,999 BDT':'LMC'
7. 'Less than 30,000 BDT':'LOI'.

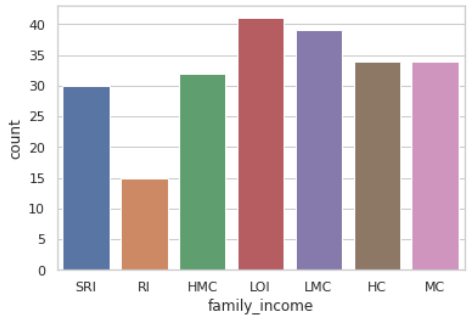

Figure 7: Family Income

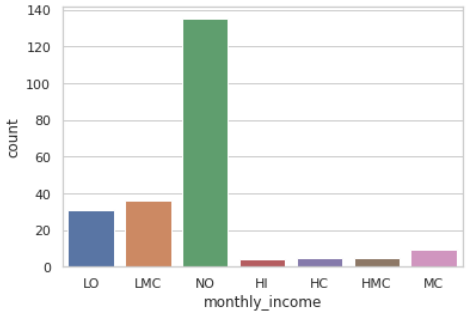

Figure 8: Own Monthly Income

In the case of Figure 8, which represents the monthly income for oneself, we have used the following scale:

1. 'More than 1,00,000 BDT':'HI',
2. '70,000 BDT - 99,999BDT':'HC',
3. '50,000BDT - 69,999BDT':'HMC',
4. '30,000BDT - 49,999BDT':'MC'
5. '50,000 - 69,999 BDT':'MC',
6. '30,000 - 49,999 BDT':'LMC'
7. '10,000BDT - 29,999BDT':'LMC',
8. 'I do not have any income at the moment':'NO',

Here, we can also see that there are respondents from various income classes in our dataset but a large number of them do not earn themselves.

Getting infected by a disease like COVID19 definitely affects one psychologically. So, we wanted that data to predict the severity of one's anxiety. Figure 9 represents information about whether the respondent himself/herself has contracted COVID19 and Figure 10 shows data regarding whether COVID19 has infected any of the family members.

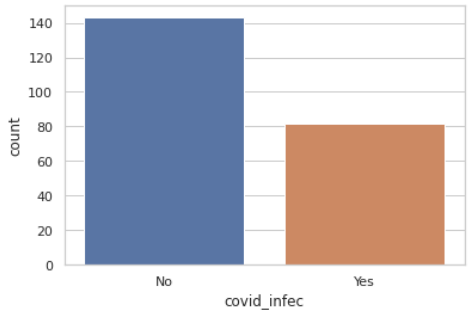

Figure 9: Infected with COVID19

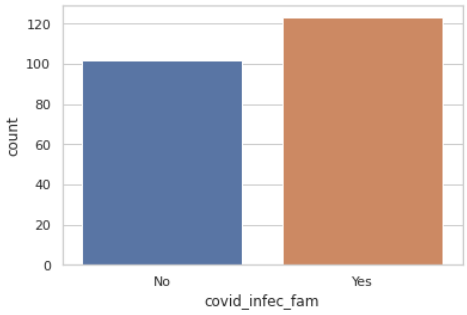

Figure 10: Any Family Member Infected with COVID19

We also wanted to see how many of the respondents knowingly experienced anxiety or other psychological issues during the pandemic. Figure 11 is used to show that representation and figure 12 shows how many of them had sought professional help. It is clear that even after experiencing psychological distress, most people refuse to get mental health counseling or similar treatments.

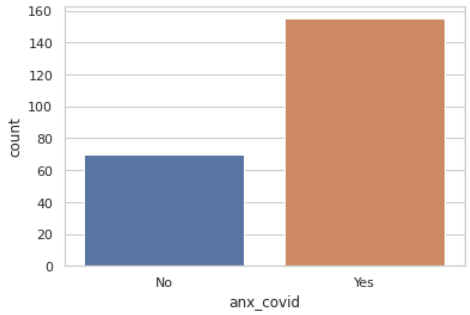

Figure 11: Experienced Psychological Distress

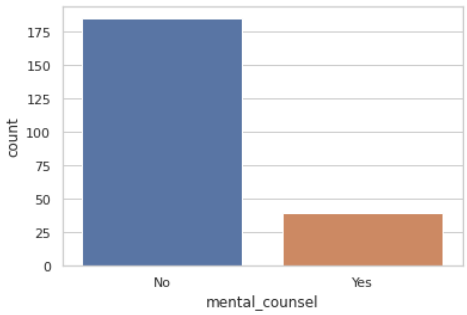

Figure 12: Received Mental Health Treatment

There are seven questions in GAD-7 assessment and the score of each questions are summed to get the final result. We have followed the similar process to determine the level of anxiety. After adding the scores of all the questions, we have stored them in a different column. This column alone is used to understand the anxiety level of each respondent. We have also assigned labels to each final score following the GAD-7 scale(FIgure 2). To keep our dataset clean, we have dropped the individual scores of each questions as we do not need them anymore.

4.3 Algorithms

There are various types of algorithms for different problems in machine learning. The problem that we are trying to solve is a classification problem meaning the model will classify whether there is a specific problem or not. In this project, we have used the following algorithms, K Nearest Neighbours, Decision Tree, and Logistic Regression. We have used statistical techniques such as SMOTE, and cross-validation. SMOTE is applied to increase the number of cases in our dataset. Cross-validation is a method to evaluate and compare learning algorithms after dividing data into multiple segments. We have also used a particular cross-validation technique, stratified K-fold sampling,

to split the dataset and keep the ratio of target classes the same throughout the whole dataset. Another implemented cross-validation technique is shuffle split which randomly samples the entire dataset during each iteration to generate training and a test set.

K-Nearest Neighbours works by observing similarities between existing data and new data. This model stores all the data and when new data is assigned, it analyzes that data by comparing it to each of the current data. When the KNN model finds data mostly similar to a particular existing data, it assigns a similar label to the new data. KNN is particularly useful for classification problems. Thus, we have implemented this model in our project.

Decision Tree is one of the most popular models used for classification problems. It is called a tree because it works with different nodes and each of these nodes denotes an attribute. There are also branches that represent the outcome of a node. Using this tree-like feature can lead to a decision. This model has the ability to handle high-dimensional data. It is also proven to have good accuracy in classification problems.

Logistic regression is used to predict a binary outcome based on the analysis of a dataset. It takes into account the correlation between different features of the dataset. It is one of the most commonly used algorithms in classification problems. As in this project, we are working on a classification problem, this model is suitable.

5 Analysis

According to the GAD-7 method, we have worked with four different levels of anxiety on this project and they are: Minimal, Mild, Moderate, and Severe. The model has predicted certain levels using the collected data. These four levels have been maintained regardless of the algorithm used to train the model.

After applying KNN model, at first, it was able to score a test accuracy of 40%. However, after modifying our data using SMOTE and other cross validation techniques, KNN managed to score accuracy of 53.4%. In this model, we have used maximum 30 neighbours to train the model and it has shown the highest accuracy when 15 neighbours were considered.

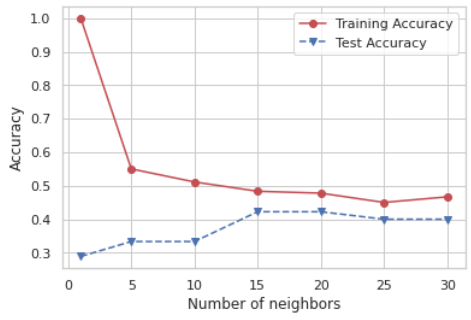

Figure 13: Initial KNN Accuracy

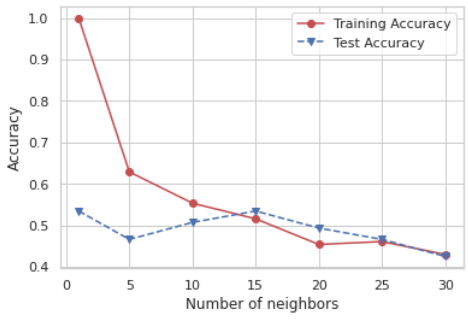

Figure 14: Final KNN Accuracy

Initially, on our dataset, decision tree produced accuracy score of 35%. After modifying our dataset, this score had risen to 54.7%. We have used a maximum depth of 50 in this model. The algorithm has shown the highest accuracy at depth of 8.

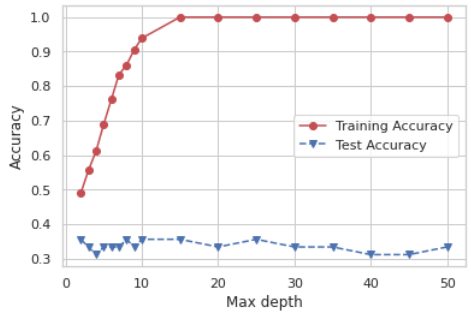

Figure 15: Initial Decision Tree Accuracy

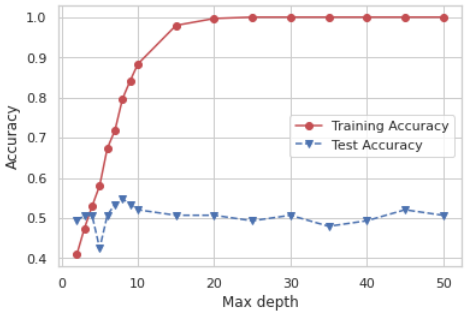

Figure 16: Final Decision Tree Accuracy

Logistic regression has produced an accuracy score of 50.5% on our dataset. Observing the confusion matrix(Figure 17), it is evident that our trained model is more accurate in predicting 'Minimal' and 'Severe' cases than other anxiety levels.

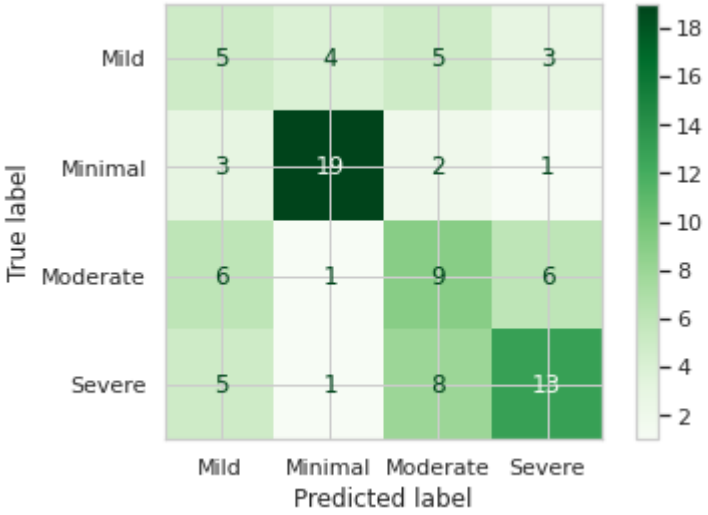

Figure 17: Confusion Matrix

We have compiled the accuracies of all the algorithms in Figure 18. We have used three of the most common algorithms used in classification problems. It is evident that the decision tree algorithm has worked better than other algorithms for our dataset.

Classifier Name	SMOTE	Label	Accuracy
KNN	Before SMOTE	Minimal	0.40
		Mild	
		Moderate	
		Severe	
	After SMOTE	Minimal	0.534
		Mild	
		Moderate	
		Severe	
Decision Tree	Before SMOTE	Minimal	0.35
		Mild	
		Moderate	
		Severe	
	After SMOTE	Minimal	0.547
		Mild	
		Moderate	
		Severe	
Logistic Regression	Before and after SMOTE	Minimal	0.50
		Mild	
		Moderate	
		Severe	

Figure 18: Results

6 Conclusion

6.1 Summary

The main goal of this project was to build a machine learning model which is able to predict the general anxiety level based on their lifestyle during the pandemic. We have collected data from 230 people on their personal lifestyles and mental health. Analyzing the collected data, we have used various machine learning algorithms to build the most efficient model for this purpose. In our study, we have found that the Decision Tree has accurately predicted the outcome most of the time. The accuracy of the decision tree algorithm was 54.7%. KNN and logistic regression managed to acquire accuracy scores of 53.4% and 50.5% respectively. Thus, for our dataset, decision tree has achieved higher accuracy than other algorithms.

6.2 Limitations

The major limitation of this project is the lack of quality data. With a larger number of data, this project would have performed better. There is also a higher number of respondents who are from a similar background and this has caused a certain level of bias in the dataset. Professionals can carry out future studies on this similar topic keeping in mind the mentioned lackings.

7 References

[1] Ana Bártolo, Sara Monteiro, and Anabela Pereira. "Factor structure and construct validity of the Generalized Anxiety Disorder 7-item (GAD-7) among Portuguese college students". In: *Cadernos de saude publica* 33 (2017).

[2] Yang Eun Kim and Boram Lee. "The psychometric properties of the patient health questionnaire-9 in a sample of Korean university students". In: *Psychiatry investigation* 16.12 (2019), p. 904.

[3] Bernd Löwe et al. "Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population". In: *Medical care* (2008), pp. 266–274.

[4] Reza Omani-Samani et al. "Prevalence of generalized anxiety disorder and its related factors among infertile patients in Iran: a cross-sectional study". In: *Health and Quality of Life Outcomes* 16.1 (2018), pp. 1–5.

[5] Tracy A Prout et al. "Identifying predictors of psychological distress during COVID-19: a machine learning approach". In: *Frontiers in Psychology* 11 (2020), p. 586202.

[6] Arkaprabha Sau and Ishita Bhakta. "Predicting anxiety and depression in elderly patients using machine learning technology". In: *Healthcare Technology Letters* 4.6 (2017), pp. 238–243.

[7] Robert L Spitzer et al. "A brief measure for assessing generalized anxiety disorder: the GAD-7". In: *Archives of internal medicine* 166.10 (2006), pp. 1092–1097.

[8] Qiu-Yue Zhong et al. "Diagnostic validity of the generalized anxiety disorder-7 (GAD-7) among pregnant women". In: *PloS one* 10.4 (2015), e0125096.

8 Appendix

Material

Github Repository

Terms

GAD = General Anxiety Disorder

SMOTE = Synthetic Minority Oversampling Technique

Cronbach's α = Cronbach's alpha is a way of assessing reliability by comparing the amount of shared variance, or covariance, among the items making up an instrument to the amount of overall variance.

Confusion matrix = A confusion matrix is a table that is used to define the performance of a classification algorithm. It represent counts from predicted and actual values.

Data Collection Form

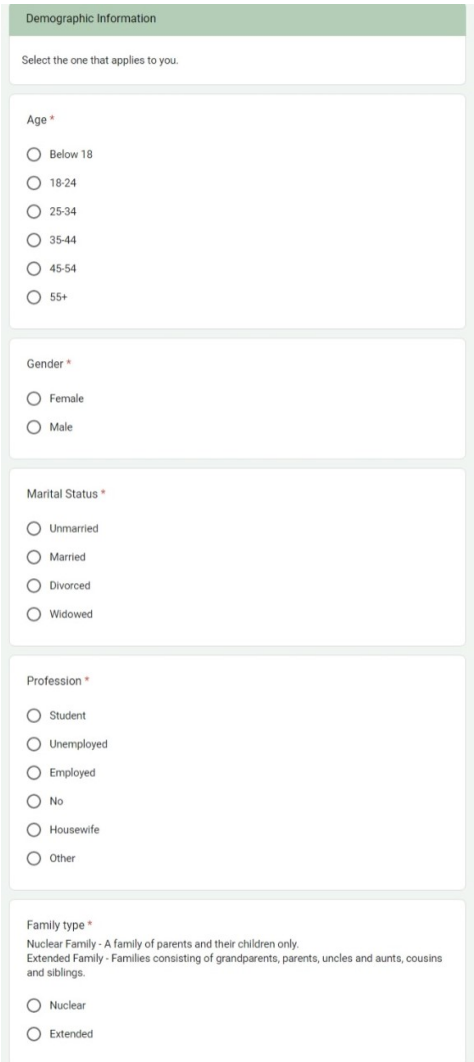

Figure 19: Demographic Information

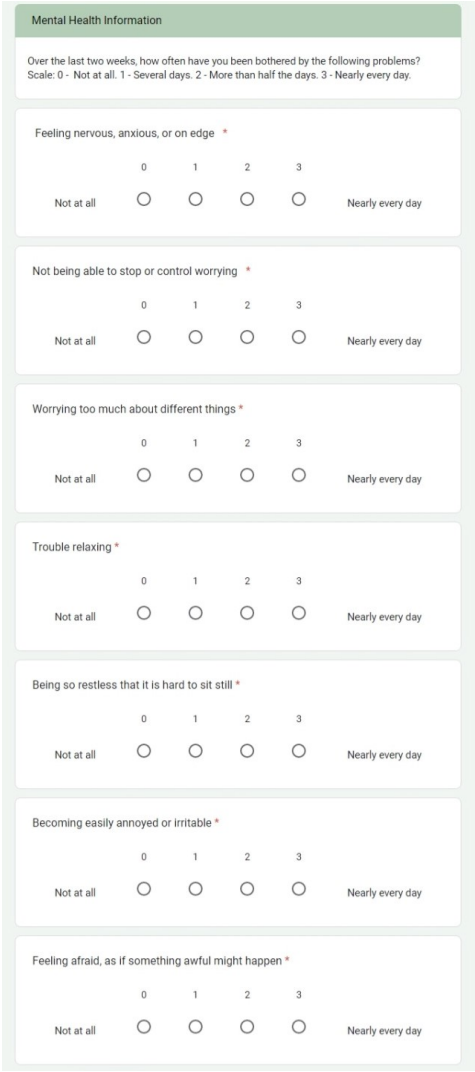

Figure 20: GAD-7 Questions

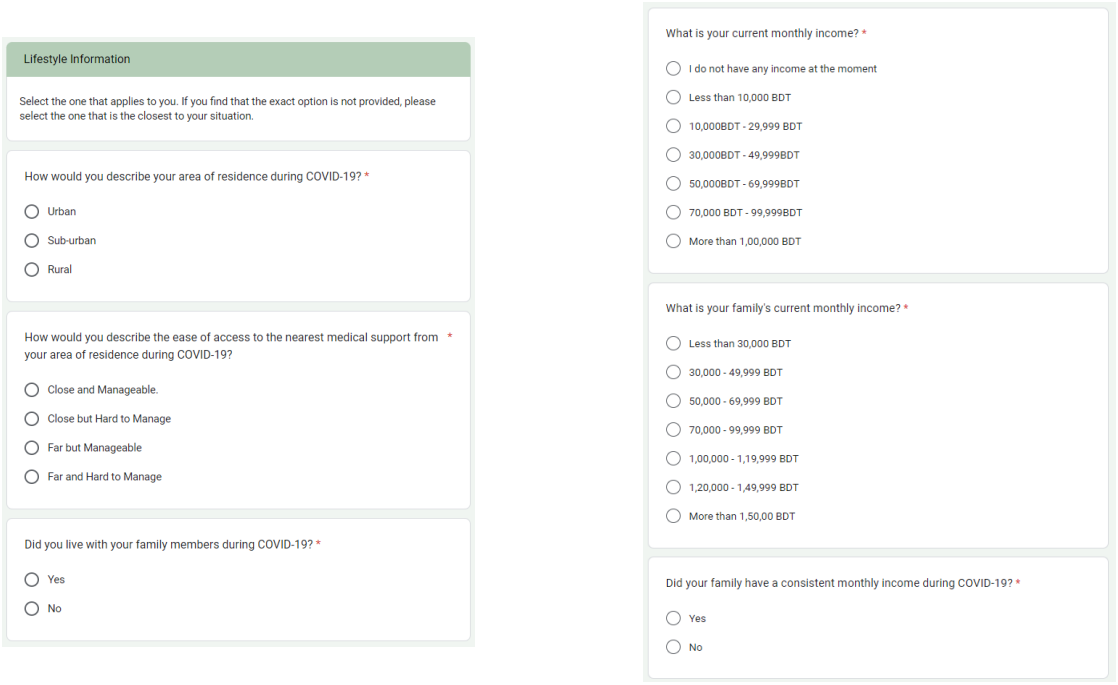

Figure 21: Lifestyle Information

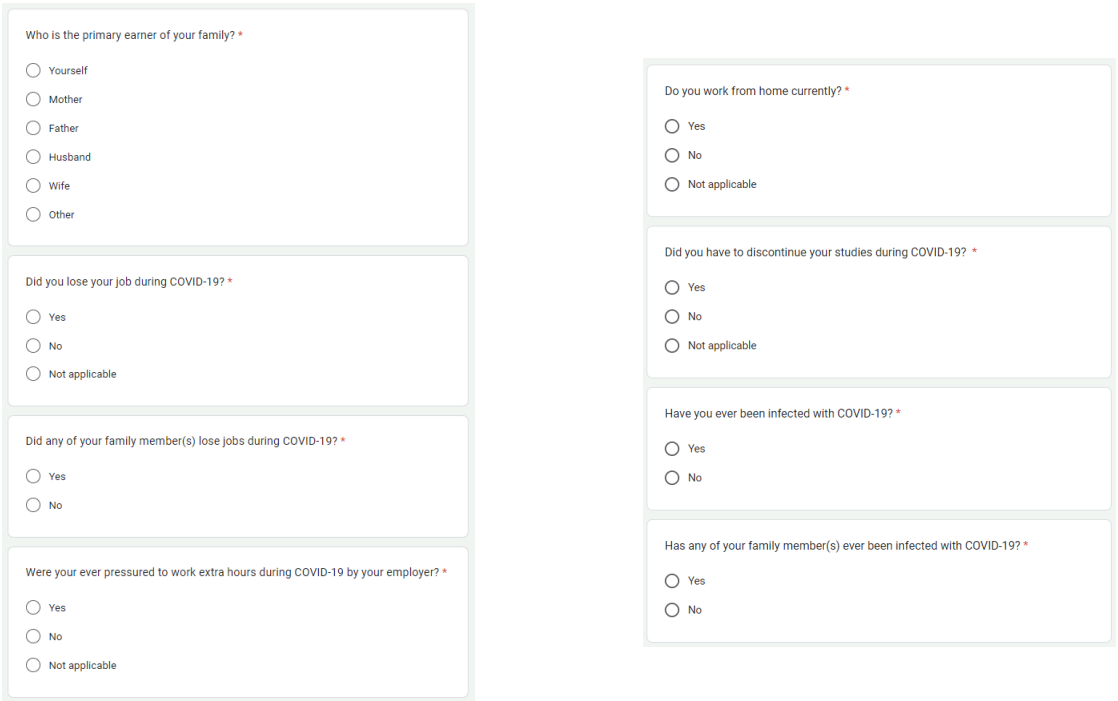

Figure 22: Lifestyle Information

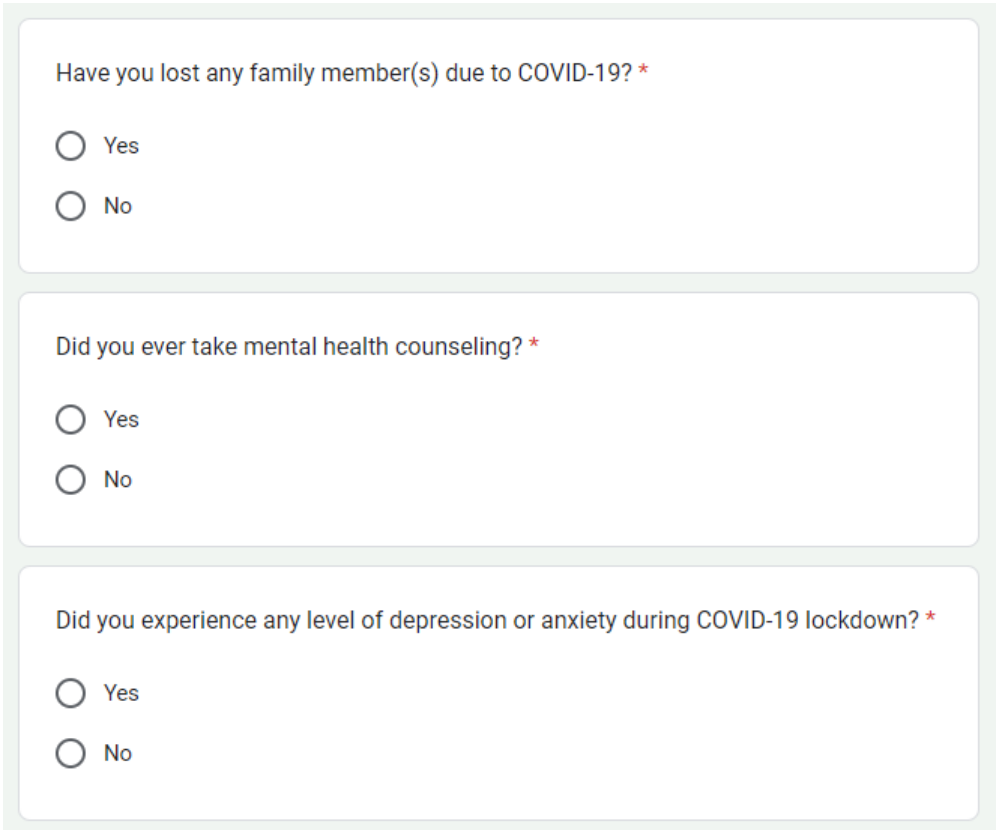

Figure 23: Lifestyle Information