A machine learning approach to predicting anxiety and depression levels amongst Bangladeshi students following the reopening of educational institutions

CSE498R (Directed Research) report submitted in partial fulfillment of the requirements for the degree

of

Bachelor of Science in Computer Science and Engineering

by

Erfan Mostafiz	Md Toufique Husein	Abdullah Al Mahfug	Adil Bin Mohammad Himal
ID: 1912734042	ID: 1921750642	ID: 1821861642	ID: 1722175042
erfan.mostafiz@northsouth.edu	toufique.husein@northsouth.edu	abdullah.mahfug@northsouth.edu	adil.himel@northsouth.edu

Under the Supervision of:

Dr. Sifat Momen

Associate Professor

Department of Electrical & Computer Engineering

North South University

Summer 2022

DECLARATION

Project Title	A machine learning approach to predicting anxiety and depression levels amongst Bangladeshi students following the reopening of educational institutions
Authors	*Erfan Mostafiz, Md Toufique Husein, Abdullah Al Mahfug, Adil Bin Mohammad Himal*
Student IDs	1912734042, 1921750642, 1821861642, 1722175042
Supervisor	Dr. Sifat Momen

This report is prepared as a requirement of the CSE498R Directed Research course. We declare that this CSE498R report entitled *A machine learning approach to predicting anxiety and depression levels amongst Bangladeshi students following the reopening of educational institutions* has not been accepted for any degree and is not concurrently submitted in candidature of any other degree. We would like to request you to accept this report as a partial fulfillment of Bachelor of Science in Computer Science and Engineering degree under Electrical and Computer Engineering Department of North South University.

Erfan Mostafiz (ID: 1912734042)

Department of Electrical & Computer Engineering, North South University

Md Toufique Husein (ID:1921750642)

Department of Electrical & Computer Engineering, North South University

Abdullah Al Mahfug (ID: 1821861642)

Department of Electrical & Computer Engineering, North South University

Adil Bin Mohammad Himal (ID: 1722175042)

Department of Electrical & Computer Engineering, North South University

Date: 11th September, 2022

Department of Electrical & Computer Engineering,
North South University
Bashundhara, Dhaka-1229, Bangladesh

APPROVAL

This is to certify that the CSE498R report entitled **A machine learning approach to predicting anxiety and depression levels amongst Bangladeshi students following the reopening of educational institutions**, submitted by Erfan Mostafiz (Student ID: 1912734042), Md Toufique Husein (Student ID: 1921750642), Abdullah Al Mahfug (Student ID: 1821861642) and Adil Bin Mohammad Himal (Student ID: 1722175042) are undergraduate students of the **Department of Electrical & Computer Engineering**, North South University. This report partially fulfils the requirements for the degree of Bachelor of *Science in Computer Science and Engineering* on September 11, 2022, and has been accepted as satisfactory.

Dr. Sifat Momen
Associate Professor

Department of Electrical & Computer Engineering
North South University

Dr. Rajesh Palit
Professor & Chair
Department of Electrical & Computer Engineering
North South University

Place: Dhaka, Bangladesh

ACKNOWLEDGEMENTS

- We dedicate this work to the numerous people around the world who are suffering from depression and anxiety, especially in the post COVID19 era.

- We acknowledge the continuous guidance we received from our supervisor Dr. Sifat Momen and for his prolonged efforts in helping us gather the data from the students which was needed in our work.

Erfan Mostafiz, Md Toufique Husein, Abdullah Al Mahfug and *Adil Bin Mohammad Himal*

North South University

Date: September 11th, 2022

Abstract: Bangladesh kept its schools, colleges and universities closed, either partially or fully, for a period of over 82 weeks, the longest period of closure of educational institutions in the world. This negatively hampered the mental health of many students of the country and many of them faced difficulties in coping with the new normal as the institutions were reopened after 2 years. During the time this paper is written, Bangladesh has no such public dataset that contains the mental health condition of the students based on their social, lifestyle and health features, following the reopening of the educational institutions. In this research, we look in the problem of predicting anxiety level and depression level amongst students of Bangladesh after the reopening of schools, colleges and universities. A survey is conducted in which students from 30 different institutions from many parts of Bangladesh took part in, ranging from students of urban areas to rural towns and from a variety of socio-economic demographics. The initial dataset is then preprocessed to make it ready for supervised machine learning algorithms to work with. Our two target variables which are being predicted are *Anxiety Level* and *Depression Level*. Exploratory Data Analysis (EDA) has been performed on the data which produced several important trends related to reasons why students have severe to moderate anxiety or severe to moderate depression. Following that, various machine learning classification algorithms have been applied over the preprocessed dataset, including Support Vector Machine (SVM), Light Gradient Boosting Machine (LGBM), K-Nearest Neighbor (KNN), Logistic Regression, Random Forest and Decision Tree. After exploring the hyperparameters used in each classifier using GridSearchCV, the best hyperparameters have been used to train the models of the six classifiers two times, once for predicting the *Anxiety Level* target variable and once for predicting the *Depression Level* target. After evaluating the performance of each model using several evaluation metrics like accuracy, precision, recall and f1-score, KNN gave the best accuracy for predicting both the target variables of Anxiety Level and Depression Level.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Emotion is one of the key components that affect people directly in their day-to-day lives. In this modern era, many people around the world suffer from anxiety and depression, especially students in their studies. During the Covid-19 pandemic, which shook the world and the people to its core, the infection spread to every country and territory around the globe. Because of this pandemic, people had to stay quarantined. People were worried about the virus and governments all around the world kept their schools closed. According to a data by UNICEF, in total, 131 million students from 11 different countries have missed their three-quarter in-person learning from March 2020 to September 2021. Among them, Bangladesh has the greatest number of students who missed all in-person classes during this period, which amounts to 36.8 million students of the country or 28.1% of the total students who missed in-person classes [1].

This is because Bangladesh's school closure lasted for over 82 weeks since the start of the Covid19 pandemic and has been the longest closure of educational institutions (fully or partially) in the world due to Covid19, according to Unesco [2]. This surely had an impact on the students' mental health due to this long period of closure. Although online classes were conducted, interactions between fellow students and teacher-student relationships have remained void due to lack of physical presence in the classroom. In our research, we analyzed the mental health conditions of the students following the re-opening schools after almost 2years, with the help of machine learning techniques.

1.1 Collecting data of Mental Health of Bangladeshi students after the reopening

In our research, with the help of machine learning, we analyzed the mental health of the students based on anxiety and depression level as the educational institutions re-opened after almost 2 years of the start of the pandemic. The GAD-7 scale [3] was used to access the students' anxiety level and the PHQ-9 scale [4] was used to access the students' depression level. Before our work, there has been no publicly available data to analyze anxiety or depression level of students in Bangladesh after the reopening of education institutions. Many students have been experiencing these negative emotions after the reopening of their institutions since many have faced problems in adapting to the life changes after doing online classes for over 2 years. Therefore, to collect data about the students' mental health, we devised a 52-set questionnaire that asked important questions relating to mental health and lifestyle methods before and after the reopening of educational institutions. A survey was conducted based on these 52 questions and students from all over the country from 30 different institutions, including school, college and universities, participated in the survey.

1.2 Applicability of Machine Learning in Predicting the Anxiety and Depression Level of students

After collecting the dataset, it was preprocessed for machine learning algorithms to run effectively. Several important trends and distributions have been found pertaining to anxiety and depression level of students. These trends are shown in the Data Visualization section in the form of charts and plots. 6 different supervised machine learning classification algorithms have been performed on the dataset

of which K-Nearest Neighbor gave the best accuracy followed by Support Vector Classifier (SVC) and Random Forest.

1.3 Research Goal

The goals of this research are summarized as follows:

1 To conduct a survey in order to collect dataset relating to mental health of Bangladeshi students after the reopening of their educational institutes.
2 To identify prominent trends pertaining to anxiety and depression level in students based on several life factors.
3 To predict the anxiety level and depression level of students using supervised machine learning techniques.

Our analysis follows the steps shown below in the flowchart of Figure 1.1.

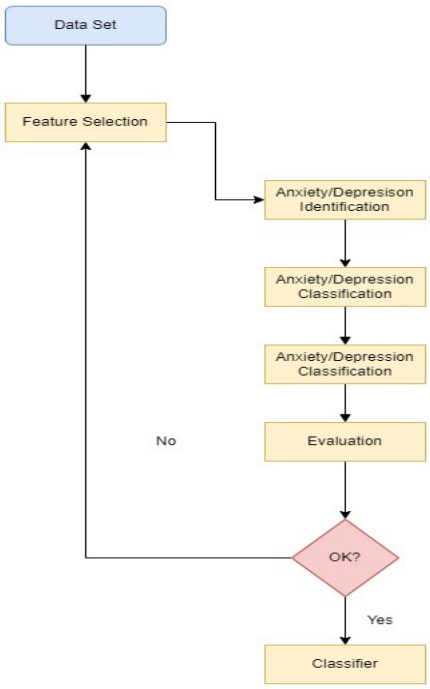


Figure 1.1 - Steps in Anxiety/Depression Analysis.

The rest of this paper is organized as follows: Chapter 2 analyzes and reviews the related works in this field of study. Chapter 3 is dedicated to elaborately explaining all the steps of our research methodology. Chapter 4 shows the results and findings of our work. Finally, the paper has been concluded in Chapter 5.

Chapter 2

Related Works

Jingyi et. al. [5] set out to conduct a cross-sectional study among primary school students in Songjiang District of Shanghai and Taizhou of Zhejiang Province, with varying levels of educational and economic development in East China between June 26 and July 6, 2020, when primary schools in these areas were reopened. Three classes from each grade (Grades 1-5) were randomly chosen in each of the four primary schools (Key Schools and Non-Key Schools) that were randomly sampled from each region. After removing the 204 participants who declined to take part in the study, a total of 6400 students from these classes were included. Each student's primary caregiver was requested to respond to a questionnaire that was distributed via an online platform in China in order to report information about their child. The final subscale was used to calculate the prosocial behavior scores, with lower scores indicating less prosocial behavior. According to the total difficulties subscale scores ("Normal," score 0-13; "Borderline," score 14-16; "Moderately abnormal," score 17-19; and "Prominently abnormal," score 20-40), there are four levels of total difficulty. After the start of the new school year, they found that among students in primary schools, the prevalence of borderline to noticeably abnormal scores was 12.46% for overall difficulties and 45.12% for prosocial behavior.

Ziyuan et. al. [6] aimed to assess the psychological effects of COVID-19 following the start of classes and investigate, using machine learning, the variables that affect students' levels of anxiety and depression. Using the Self Rating Anxiety Scale, 74 (15.5%) of the 478 valid online questionnaires gathered between September 14 and September 20 displayed symptoms of anxiety, while 155 (32.4%) displayed symptoms of depression (by Patient Health Questionnaire-9). The imbalance of the retrieved data was addressed by using the oversampling technique (SMOTE). To investigate significant influence factors, the Akaike Information Criterion (AIC) and multivariate logistic regression were used. The findings suggest that the main determinants of anxiety or depression are COVID-19 and the extent to which family economic status is influenced by it. The average area under the curve (AUC) values of the receiver operating characteristic (ROC) curves of anxiety and depression on the test set reached 0.885 and 0.806, respectively, to assess the impact of our model using 5-fold cross-validation.

Cong Zhou et. al. [7] conducted a study with the goal of examining high school students' psychological conditions following the easing of the epidemic was presented. Three high schools' demographic data, the Patient Health Questionnaire-9 (PHQ-9), the Generalized Anxiety Disorder-7 (GAD-7), the Self-Rating Scale of Sleep (SRSS), and a self-created general recent-status questionnaire were all collected via a web-based cross-sectional survey. There were 1,108 qualified questionnaires collected in total. All students reported mild to severe depressive and anxious symptoms at a rate of 27.5 and 21.3%, respectively, while 11.8% of these high-risk students experienced sleep disturbances. Their study offered insight into high school students' psychological conditions a year after the COVID-19 pandemic had been effectively contained.

Chapter 3

Methodology

The research approach used in this work is detailed in Figure 3.1.

Figure 3.1 - Flow Chart of our Research Methodology

3.1 Questionnaire Design

Questionnaires of related studies on mental health of students after the reopening of educational institutions [5] [6] were studied to get an understanding of the questions asked. These studies [5] [6]

used general questions to identify students' mental health level after the reopening of schools in China. To measure anxiety, in our study, The Generalized Anxiety Disorder scale (GAD-7) has been used [3]. The GAD-7 scale is one of the most frequently used diagnostic self-report scales for screening, diagnosis and severity assessment of anxiety disorder. It contains 7 questions, each containing 4 answers, from "0" (not at all) to "3" (nearly every day), to which the participant needs to mark one answer according to their experiences. The GAD-7 scale is detailed in Table 3.1. The anxiety scoring based on GAD-7 scale is described in Table 3.3. To measure depression, in our study, the Patient Health Questionnaire scale (PHQ-9) has been used. The PHQ-9 scale is a diagnostic tool introduced in 2001 to screen adult patients for the presence and severity of depression [4]. It contains 9 questions pertaining to measure depression among the participants, each having answers from "0" (not at all) to "3" (nearly every day). The PHQ-9 scale is detailed in Table 3.2. The depression scoring based on the PHQ-9 scale is described in Table 3.4.

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

Table 3.1 - The GAD-7 Questions

Over the last 2 weeks, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

Table 3.2 - The PHQ-9 Questions

Total Score	**Anxiety Severity**
0-4	Minimal anxiety
5-9	Mild anxiety
10-14	Moderate anxiety
15-21	Severe anxiety

Table 3.3 - GAD-7 Anxiety Scoring

Total Score	**Depression Severity**
1-4	Minimal depression
5-9	Mild depression
10-14	Moderate depression
15-19	Moderately severe depression
20-27	Severe depression

Table 3.4 – PHQ-9 Depression Scoring

A survey form was designed which contained 52 personalized questions, including the 7 questions from GAD-7 scale and 9 questions from PHQ-9 scale. The survey form can be found in [8].

3.2 Data Collection

A rigorous survey was conducted through online and offline methods in which students from schools, colleges and universities from many parts of Bangladesh took part in. In total, students from 30 different institutions of the country, aged between 15 to 28, participated in the 52 questions survey. The survey was conducted for a period of 1.5 months and a total of 183 records were collected. The initial dataset containing all the 183 instances and 56 features can be found in [9]. We then further preprocessed the dataset for our desired Machine Learning algorithms to run and details of the preprocessing part is described in the next section. The python codes that were used to preprocess, visualize, analyze and to generate machine learning prediction results can be found in [10].

3.3 Data Preprocessing

A dataset collected may not always be in a suitable format for machine learning algorithms to work with. The data pre-processing stage is crucial to ensure that machine learning algorithms can operate effectively. The total anxiety score for each participant from the 7 anxiety questions was calculated and a target variable *Anxiety_Level* was defined which contained the anxiety severity of the participants based on the GAD-7 scale. The Anxiety Severity linked to the total GAD-7 score can be found in Table-3.3. Thus the 7 anxiety features were shortened to 1 target variable of *Anxiety_Level.* The same thing was done for defining the other target variable, *Depression_Level*, which contained the depression severity of the participants based on the PHQ-9 scale. The Depression Severity linked to the total PHQ-9 score can be found in Table-3.4. Thus our 52 features data frame was reduced to 36 features including the 2 target variables *Anxiety_Level* and *Depression_Level* (Table-3.5), and 34 features which were the questions asked to the participants other than the 7 questions of GAD-7 and 9 questions of PHQ-9.

Anxiety_Level	***Depression_Level***
Minimal anxiety	Minimal depression
Mild anxiety	Mild depression
Moderate anxiety	Moderate depression
Severe anxiety	Moderately severe depression
	Severe depression

Table 3.5 - Target Variables in our model

The number of unique values in our dataset is described in Table3.6.

Type	**Feature**	**Unique Values**
Independent Features	Age	14
	Gender	2
	Institution Name	30
	Institution Type	3
	District stayed during the period of institution close	26
	Household Location	4

	Family Income Status	6
	No. of bedrooms	4
	Separate bedroom of you?	3
	Ever infected by Covid?	3
	Family member ever infected by Covid?	3
	Impact on mental health if ever infected by Covid?	5
	Lost job or business due to Covid?	3
	Impact on mental health due to deaths based on Covid	6
	Smoker?	3
	Underlying medical condition of self	19
	Underlying medical condition of family member	36
	Afraid of Covid?	4
	Ever took counselling?	3
	Ever feel the need for counselling?	3
	Depression/Anxiety during lockdown?	4
	Relationship with parents now after reopening	3
	Can't adapt to study and life changes after reopening?	4
	Lagging behind in school?	4
	Change in interaction with classmates?	4
	Performance in offline exam, compared to online	4
	Which platform (online/offline) do you interact better with teachers?	4
	Which platform (online/offline) do you understand better?	4
	Amount of computer games played during school closure	4
	Amount of computer games playing after reopening	4
	Amount of social media browsing in past week	4
	Amount of social media communication in past week	4
	Physical Exercise in the past week?	4
	Meeting with friends or relatives in the past week?	4
Target Features	Anxiety_Level	4
	Depression_Level	4

Table 3.6 - Number of Unique Values for each feature

Upon analysis no null entries were found in the dataset as it was mandatory for the participants to answer all the questions in the survey.

3.4 Exploratory Data Analysis

By exploring the preprocessed data, we identified some useful patterns, trends and distribution in the data pertaining to anxiety and depression level of Bangladeshi students after the reopening of educational institutions. Several charts, including pie chart, bar chart, violin plot, etc have been visualized with the data and are discussed below.

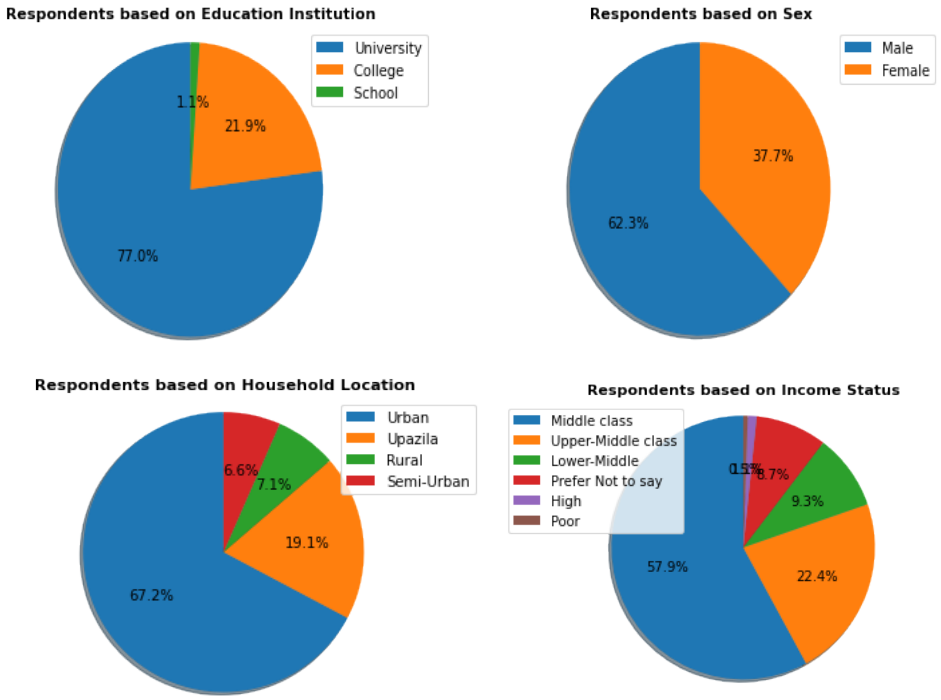


Figure 3.2 – Pie chart containing percentage of respondents based on different demographic factors

From Figure 3.2, the following statistics can be observed about the dataset:

- Most of the respondents were university students (77%) followed by college students (21.9%).
- Most of the respondents were from Urban areas (67.2%) followed by Upazila areas (19.1%).
- In terms of income status, most of the students were from Middle Class families (57.9%).

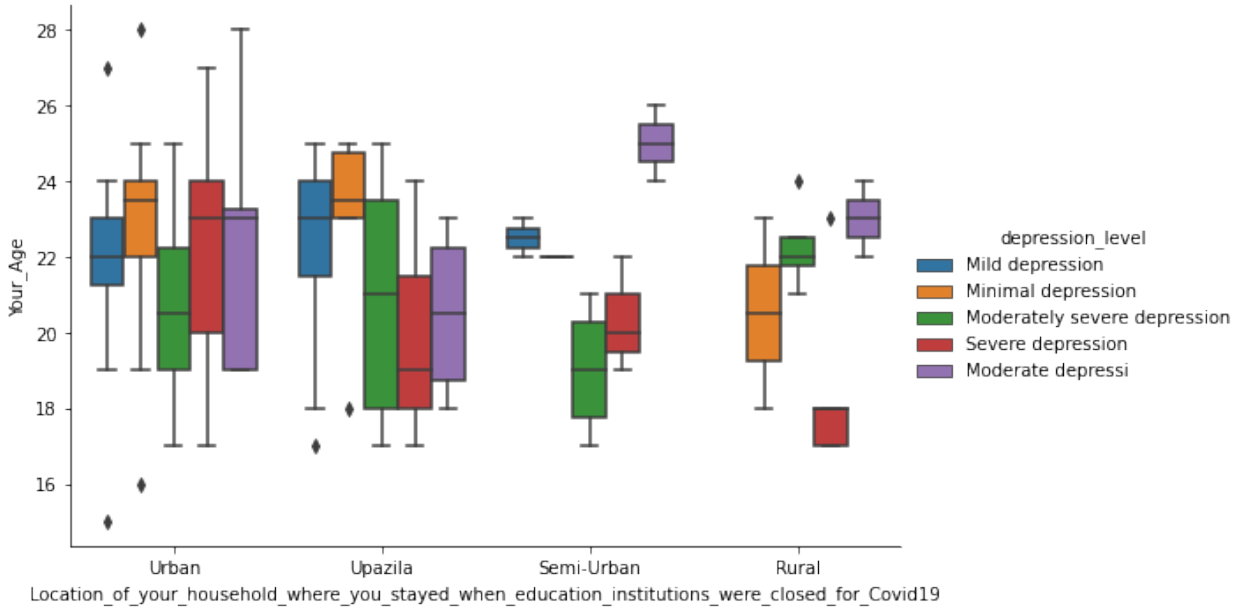


Figure 3.3 - Depression_level in different locations based on age

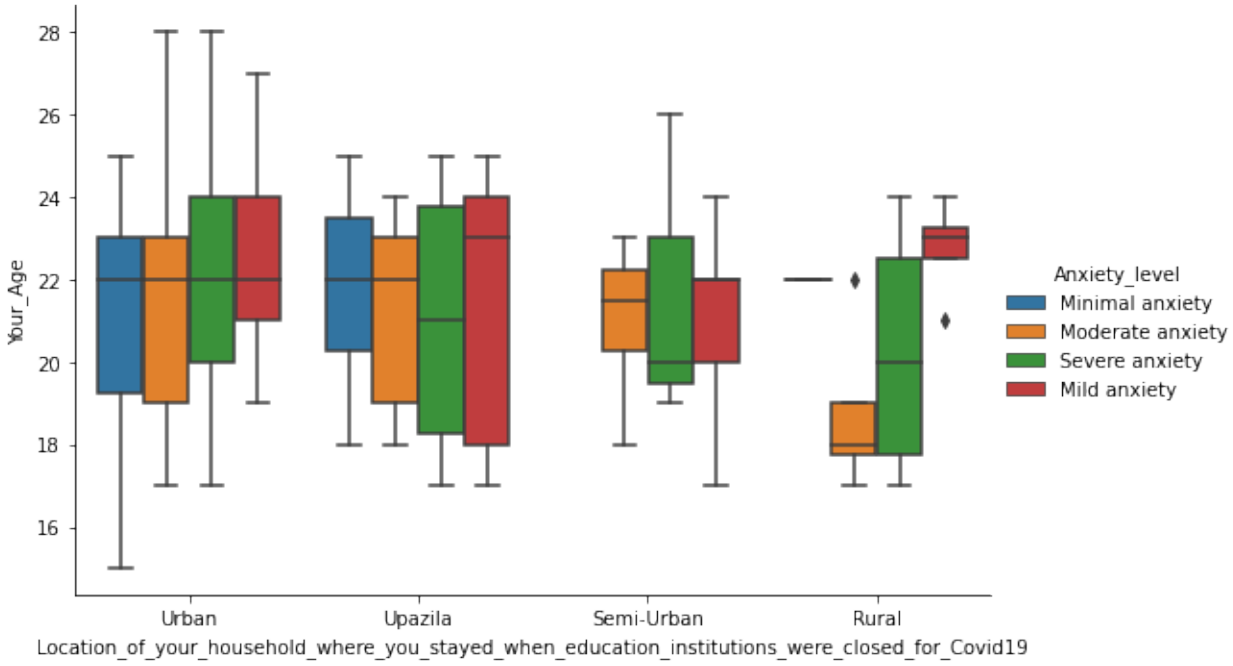


Figure 3.4 - Anxiety_level in different locations based on age

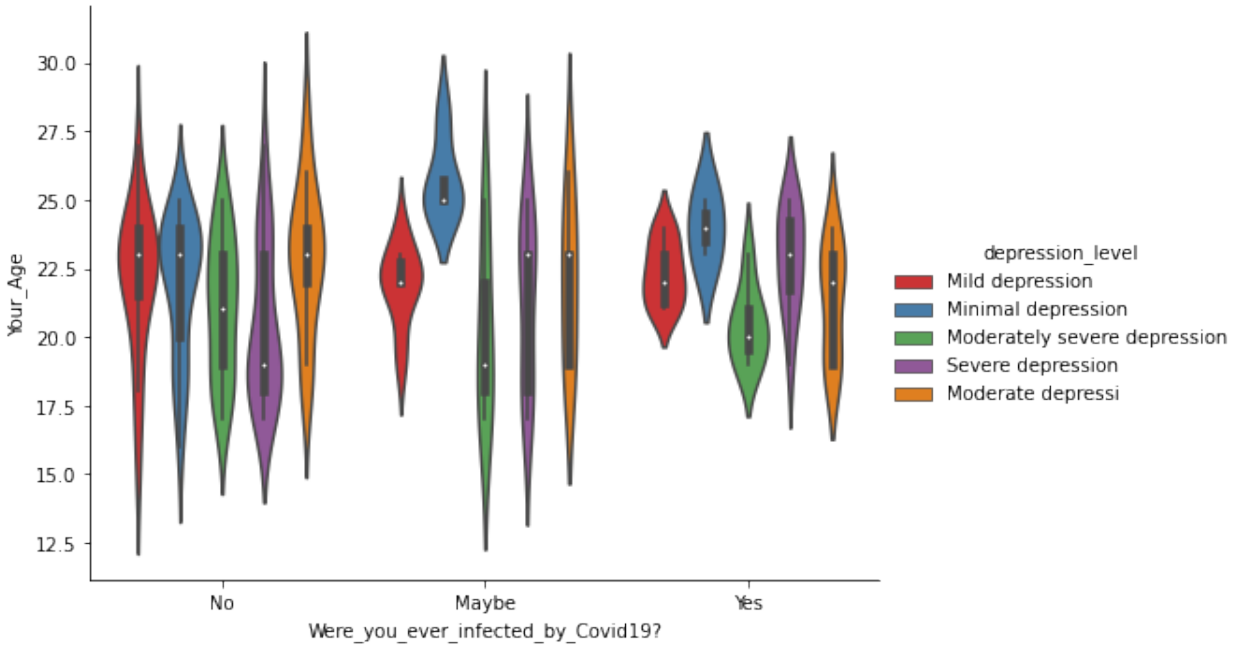


Figure 3.5 - Violin plot of depression based on if they were ever infected by Covid19 or not

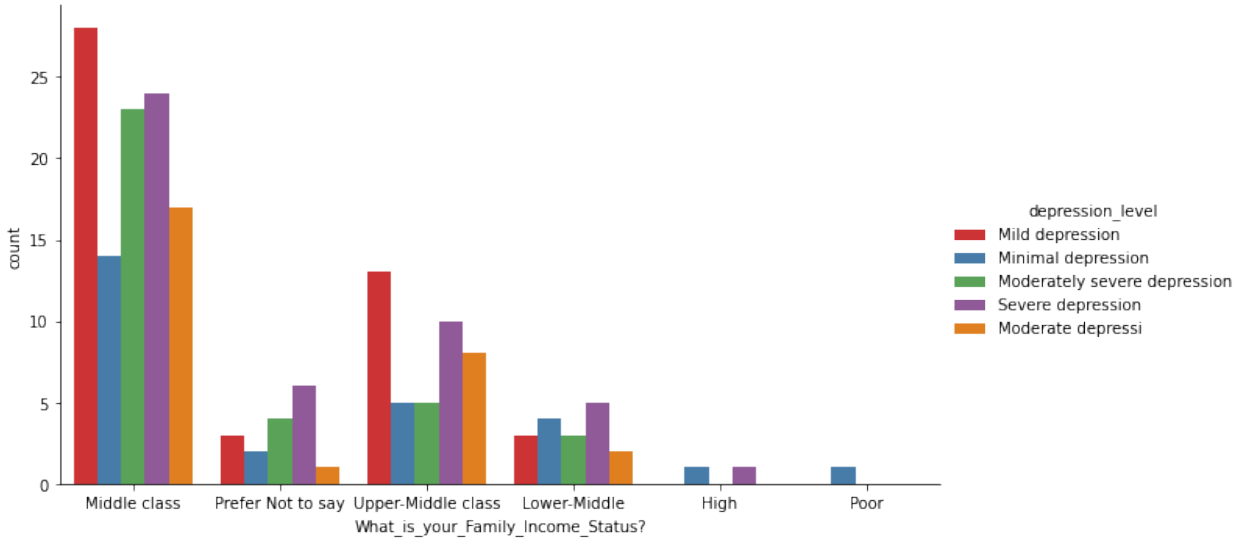


Figure 3.6 - Bar chart of depression_level according to income status

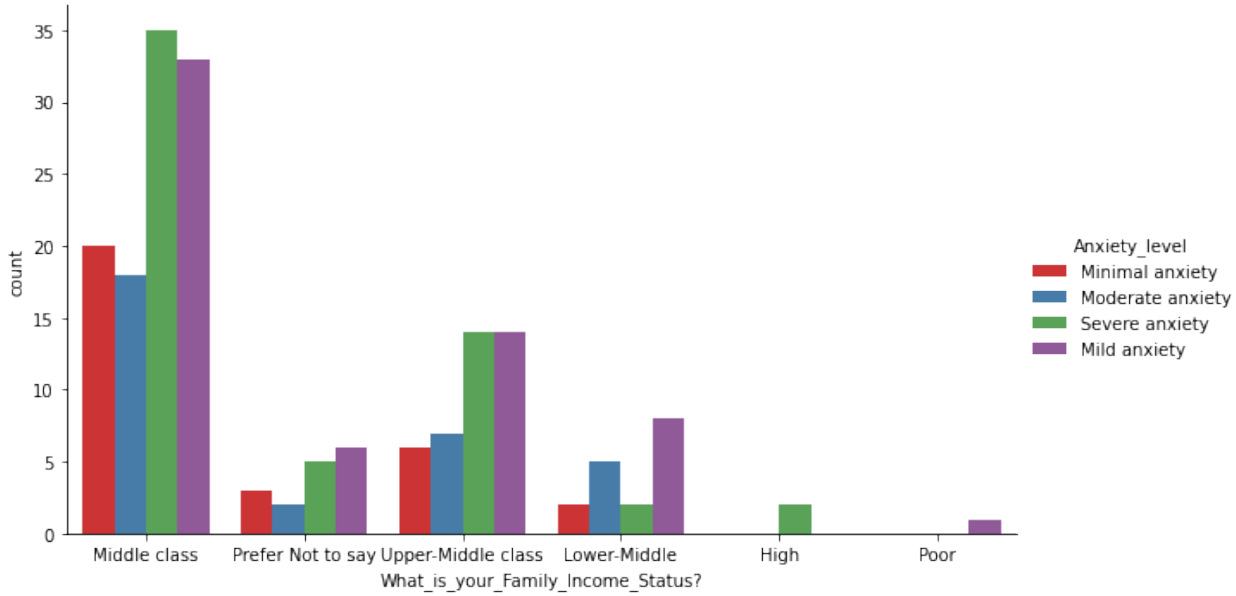
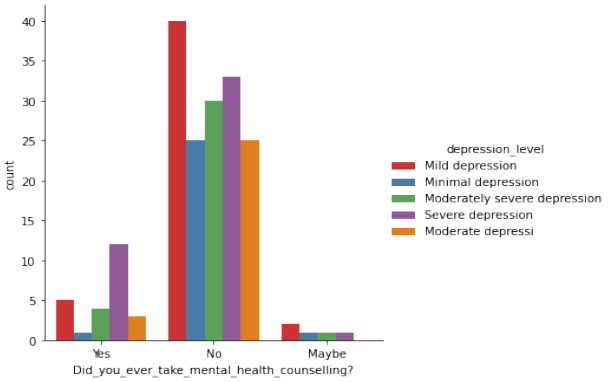
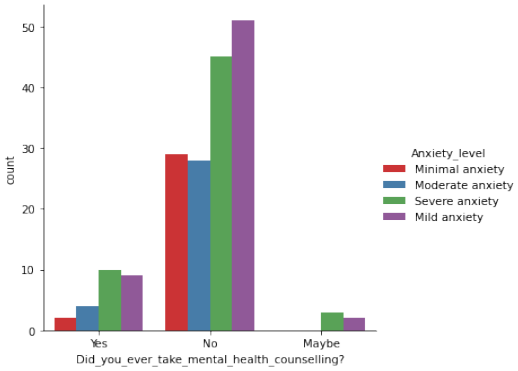
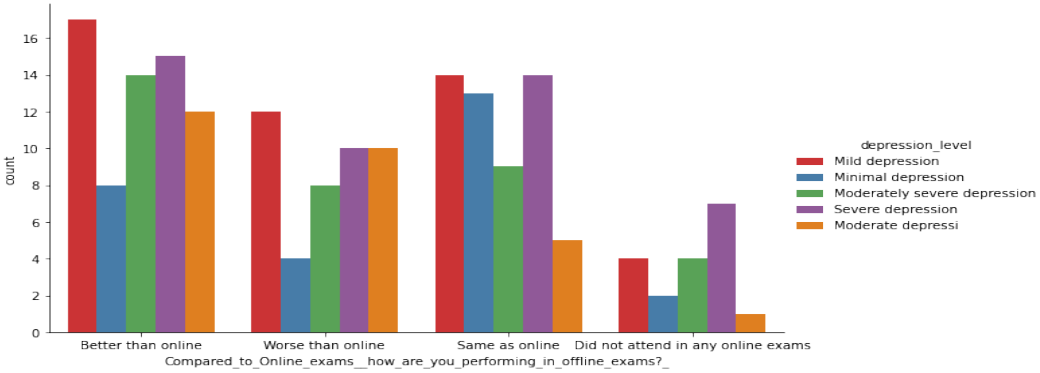
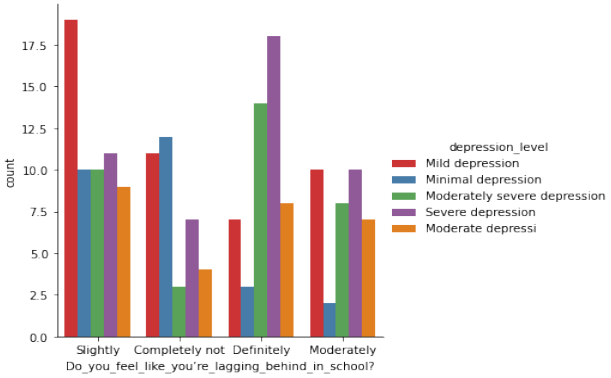


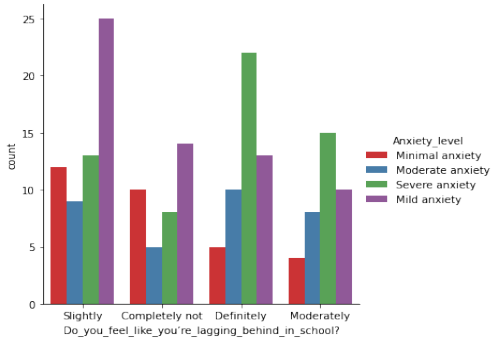
Figure 3.7 - Bar chart of anxiety_level according to income status

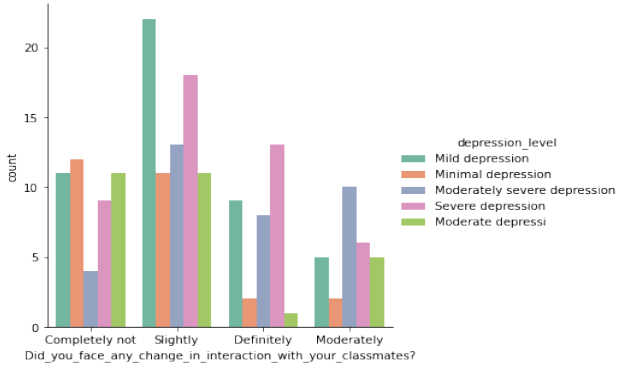

count
Yes
No
Maybe
Did_you_ever_take_mental_health_counselling?
depression_level
Mild depression
Minimal depression
Moderately severe depression
Severe depression
Moderate depressi

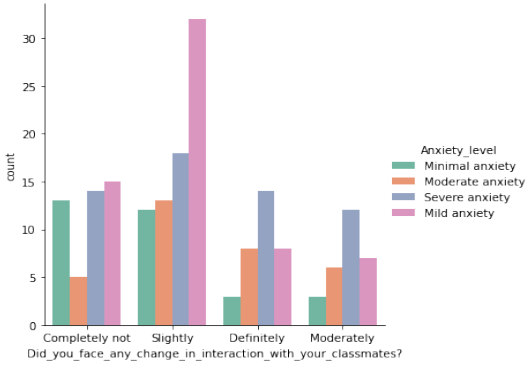

count
Yes
No
Maybe
Did_you_ever_take_mental_health_counselling?
Anxiety_level
Minimal anxiety
Moderate anxiety
Severe anxiety
Mild anxiety


count
Better than online
Worse than online
Same as online
Did not attend in any online exams
Compared_to_Online_exams__how_are_you_performing_in_offline_exams?_
depression_level
Mild depression
Minimal depression
Moderately severe depression
Severe depression
Moderate depressi


count
Slightly
Completely not
Definitely
Moderately
Do_you_feel_like_you're_lagging_behind_in_school?
depression_level
Mild depression
Minimal depression
Moderately severe depression
Severe depression
Moderate depressi


count
Slightly
Completely not
Definitely
Moderately
Do_you_feel_like_you're_lagging_behind_in_school?
Anxiety_level
Minimal anxiety
Moderate anxiety
Severe anxiety
Mild anxiety


count
Completely not
Slightly
Definitely
Moderately
Did_you_face_any_change_in_interaction_with_your_classmates?
depression_level
Mild depression
Minimal depression
Moderately severe depression
Severe depression
Moderate depressi


count
Completely not
Slightly
Definitely
Moderately
Did_you_face_any_change_in_interaction_with_your_classmates?
Anxiety_level
Minimal anxiety
Moderate anxiety
Severe anxiety
Mild anxiety

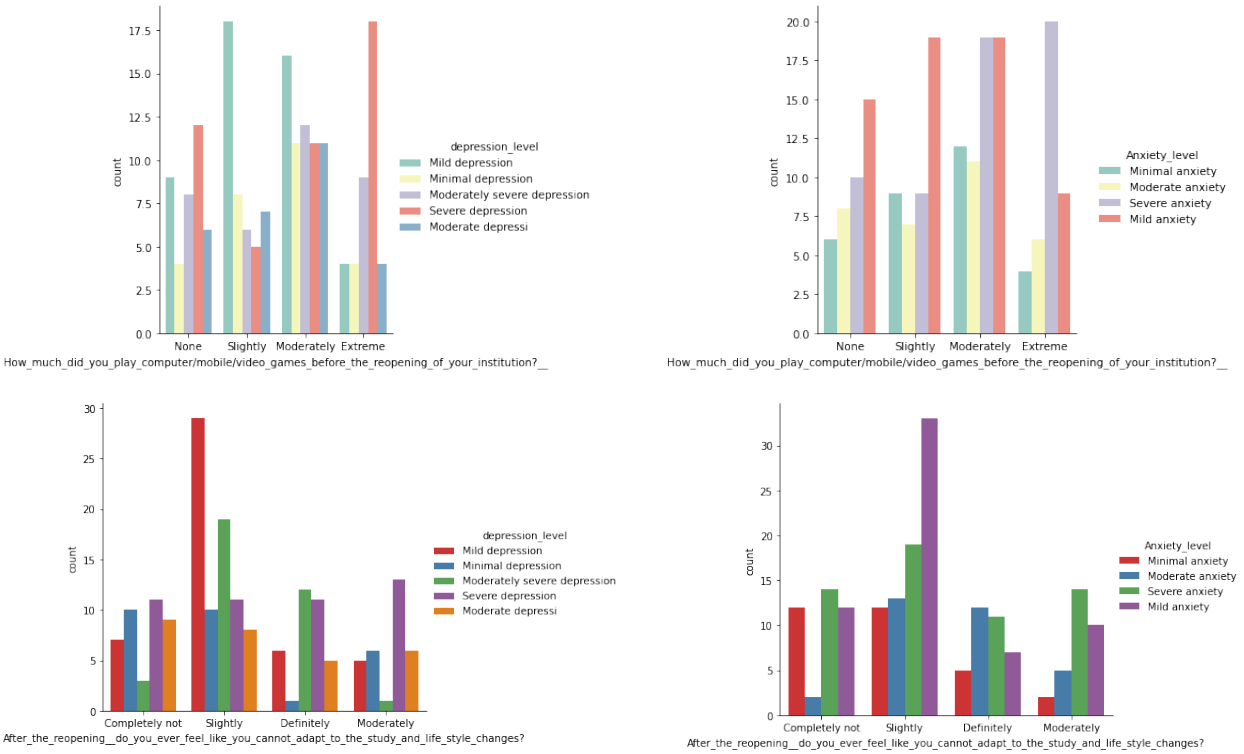


Figure 3.8 - Bar charts of depression_level and anxiety_level according to different features

Several important trends can be found amongst students' anxiety and depression level based on different features that have been outlined in Figure 3.8.

- Students who took Mental health counselling are more likely to have "severe depression" and "severe anxiety".
- Students who feel like they are lagging behind in school *completely* or *moderately* are more likely to have "severe depression" and "severe anxiety" compared to those who responded to *slightly* or *completely not.*
- Students who felt like they faced change in interaction with their classmates after the reopening of institutions *definitely* or *moderately* are more likely to have "severe" to "moderately severe" depression. They are also more likely to have "severe anxiety". Compared to them, students who feel like their interactions with their classmates has not changed after the reopening are more likely to have "minimal anxiety". This data concludes the fact that students who interact less with their classmates are more likely to have depression and anxiety.
- Students who played *extreme* levels of computer/mobile games during the closure of schools are more likely to have "severe depression" and "severe anxiety" after the reopening of schools.
- Students who feel like they cannot adapt to the life changes after the reopening of their educational institutions by *Definite* or *Moderate* amount are likely to have "severe" to "moderately severe" depression.

3.5 Data Encoding

All of the features except age had categorical variables. For most machine learning algorithms to work

effectively on the dataset, the data must be in numeric form instead of strings for categorical variables [11]. **Label Encoding** has been used to convert string data of categorical variables into numeric forms.

In Label Encoding, the categorical values are replaced with numeric values between 0 and the number of classes minus 1. If the categorical variable value contains 6 distinct classes, we use (0, 1, 2, 3, 4 and 5) to encode the data [12]. We used **LabelEncoder** [13] from scikit-learn library [14] to handle our categorical variables.

3.6 Data Scaling

In our dataset, different features have different range of values. As all features other than the *Age* feature are categorical variables and has been label encoded according to *number of classes – 1,* some features might show dominance over other features while training our model. We used the **MinMaxScaler** [15] from scikit-learn library [14] to scale our dataset and convert all the features into the range [0, 1] meaning that the minimum and maximum value of a feature is going to be 0 and 1, respectively. The formula for MinMaxScaler is mentioned in Equation 3.1.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

3.7 Data Splitting

Machine learning process requires the dataset to be split into train set and test set. By separating the test and train data, the model will not be able to know the test data and will only be trained based on the training data. This enables the models' performances to be measured accurately by evaluating them on the test set, i.e. how well the trained model performs on the unseen test set. This is because the model should be able to perform well on unseen data and that is the main purpose of building the model. We used scikit-learn library's **train_test_split** function [16] to split our dataset into 80% train set and 20% test set. Additionally, we partitioned the dataset with a fixed value for the hyper-parameter *random_state* to ensure that the results obtained are reproducible.

3.8 Feature Selection

Feature selection is the process of selecting the most important features to include in machine learning algorithms. Feature selection techniques are used to reduce the number of input variables by removing redundant or irrelevant features and limiting the feature set to those that best fit the machine learning model [17]. Certain features may be more important than others, while others may be irrelevant to predicting the anxiety or depression levels. Figure 3.9 shows a heatmap plotted to show the correlation between features.

High correlation between a pair of features indicates that the two features carry the same information, and thus it is not required to have both of them. We used the Pearson correlation co-efficient [18] to measure the strength and direction of a linear relationship between two features. Setting a threshold of 80%, we searched for the highly correlated features from the correlation matrix. We found no pair of features that had a correlation above the 80% threshold set, so no features were dropped from the dataset.

The following algorithm (Algorithm 1) is used to detect highly correlated features:

```
function CORRELATION(dataset, threshold)
    correlated_ features ← set()                    //Set of correlated features
    correlation_matrix  ← dataset.corr()    // Correlation matrix using Pearson correlation
    for every distinct pair of features do
        if difference in the correlation matrix > threshold then
            correlated_ features.insert(feature_name)
        end if
    end for
    return correlated_ f eatures
end function
```

Algorithm 1 : To Detect Highly Correlated Features

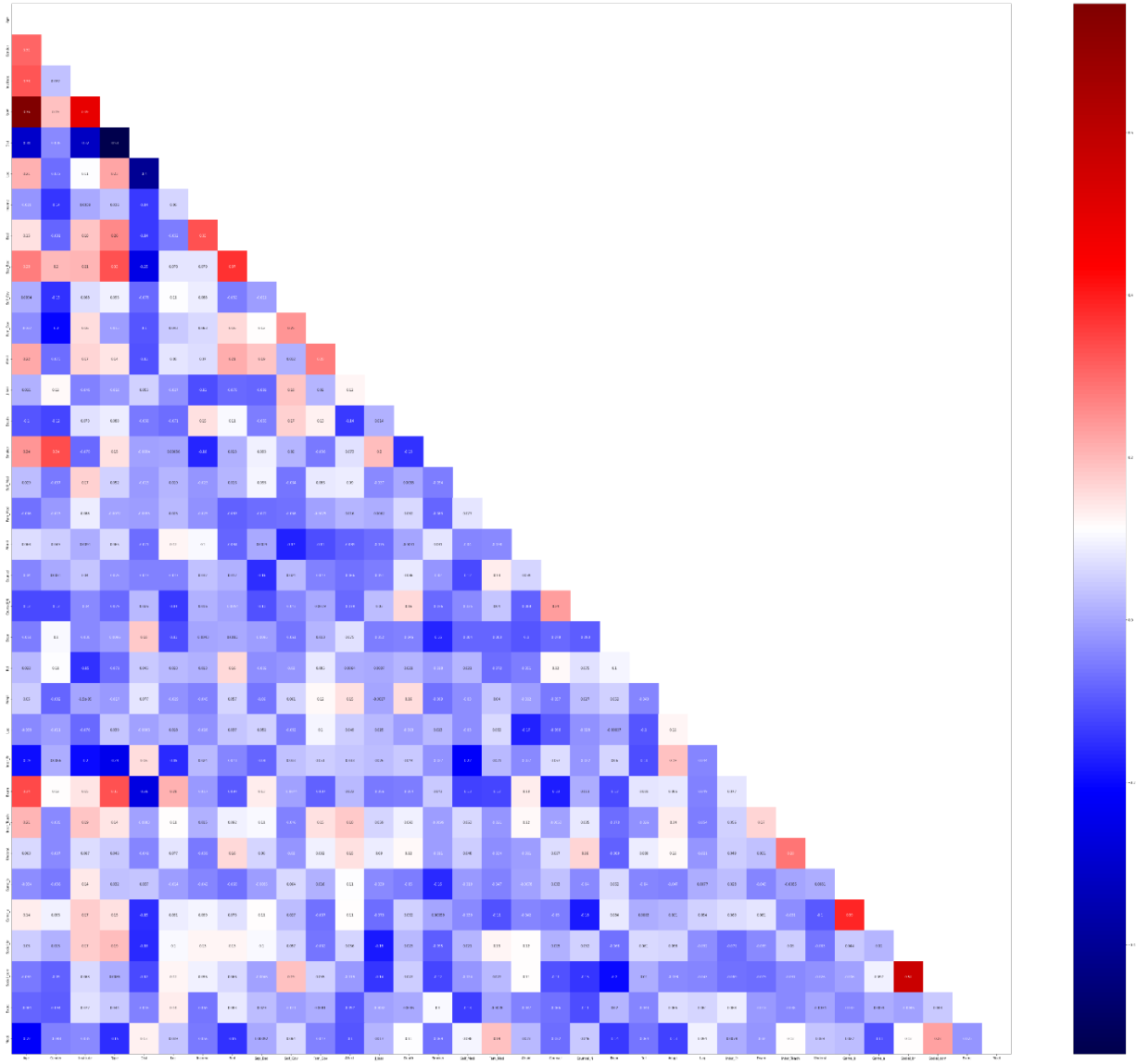


Figure 3.9 - Heatmap showing correlation between pairs of features

3.9 Classifiers used

3.9.1 Support Vector Machine (SVM)

Support Vector Machines (SVM) [19] are a set of supervised learning methods used for classification, regression, and outlier detection. There are specific types of SVMs that can be used, like support vector regression (SVR) which is an extension of support vector classification (SVC). SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.

A simple linear SVM classifier works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. This means there can be an infinite number of lines to choose from.

As our target variables, *Anxiety_level* and *Depression_level,* are categorical and contain 4 class and 5 classes respectively, so we are using Support Vector Classification (SVC) to predict the accurate class based on the features.

3.9.2 Random Forest

Random forest [20] is a supervised machine learning algorithm that can be used to solve both classification and regression problems. It is an ensemble learning method that uses multiple decision trees in order to create classification or regression model. It consists of a large number of decision trees working together as an ensemble. Each individual trees predict the value of the target class, and their predictions are combined in order to get more accurate prediction. Figure 3.10 shows the architecture diagram of Random Forest Classifier.

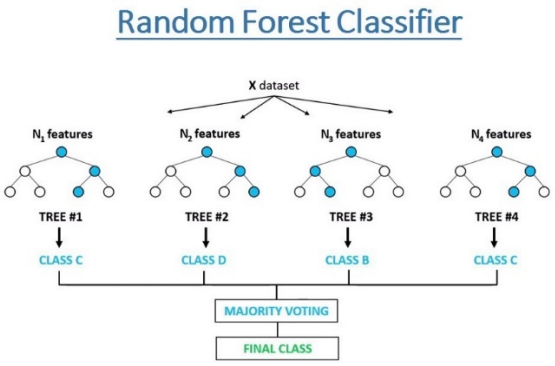


Figure 3.10 - Random Forest Classifier Tree

We have used RandomForestClassifier, imported from the scikit-learn library to predict our target variables *Anxiety_level* and *Depression_level.*

3.9.3 Logistic Regression

Logistic regression is another powerful supervised machine learning algorithm normally used for binary classification problems, when the target is categorical. It can also be used for more than two class prediction (Multinomial logistic regression). The best way to think about logistic regression is

that it is a linear regression but for classification problems. Logistic regression essentially uses a logistic function [21] defined in Equation-3.2 to model a binary output variable.

$$Logistic\ Function = \frac{e^{a+bx}}{1 + e^{a+bx}} \tag{3.2}$$

In Equation-2, x is the independent variable, a and b are parameters of the logistic model

3.9.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbor (KNN) [22] is one of the most fundamental and simplest machine learning classification algorithms based on supervised learning technique. KNN algorithm stores all the available data and classifies a new data point based on the similarity of the data. This means when a new data appears, it can be easily classified into a well suite category by using the KNN algorithm.

KNN works by measuring the Euclidean distance between sample x_i and x_j by using the following Equation-3.3

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + \left(x_{ip} - x_{lp}\right)^2} \tag{3.3}$$

In Equation-3.3, x_i is an input sample with p features, and p is the total number of features.

3.9.5 Decision Tree

Decision Tree is a supervised machine learning technique that builds regression or classification models. It uses a tree-like structure where the outcomes. All the other nodes except the leaf nodes are called decision nodes where further splits are made depending on yes/no questions. The goal of the decision tree is to create a model that can be used to predict the value of target variable by learning simple decision rules from the previous data. How a decision tree splits the data is often determined by *Entropy* or *Gini Index*. In our work, we have used Gini Index to split the data in the Decision Tree.

Equation-3.4 shows the formula for Gini Index:

$$G(S) = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{3.4}$$

where S is the subset of the training data and p_i is the probability of the class.

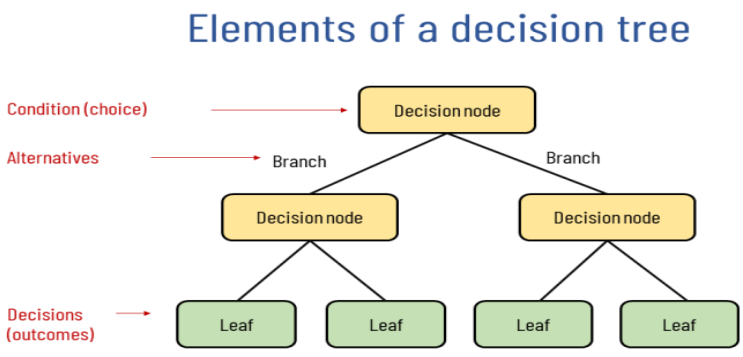


Figure 3.11 - Decision Tree Architecture

3.9.6 Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LGBM) [23], is a free and open source distributed framework for gradient boosting in the field of supervised machine learning technology, originally developed by Microsoft. It is based on decision tree algorithms and is used for ranking, classification and other machine learning tasks. Creators of the LGBM framework claim that it has the advantages of –

i) Faster training speed and higher efficiency, ii) Lower memory usage, iii) Better accuracy, iv) Support of parallel, distributed, and GPU learning, v) Capable of handling large-scale data.

3.10 Hyperparameter Tuning

While creating a machine learning model based on the classifiers or regressors used, we do not immediately know what the optimal model architecture should be for a given model, as different parameters work best for different models and different data. Parameters which define the model architecture are referred to as hyperparameters and thus this process of searching for the ideal model architecture is referred to as hyperparameter tuning [24].

We have used **GridSearchCV**, imported from scikit-learn library, to search for the best hyperparameters of a given model. Table-3.7 describes the hyperparameters we have explored for each model.

Model	Hyperparameter
Support Vector Classifier (SVC)	'C' = [1, 10, 20, 30] 'kernel' = ['rbf', 'linear']
Random Forest Classifier	'n_estimators': [1,5,10,20,50,100,200] 'random_state': [3,5,6,8] 'criterion': ['gini','entropy']
Logistic Regression	'C': [1,5,10,15,20,30] 'random_state': [3,5,8,10]
K-Neighbors Classifier (KNN)	'n_neighbors': [5,10,15,20] 'algorithm':['auto','ball_tree','kd_tree','brute'] 'weights': ['uniform','distance']
LightGBM (LGBM)	'random_state': [5, 10, 15, 20, 40, 42, 50, 60] 'max_depth' : [-1, -5, -2]
Decision Tree	'criterion': ['gini','entropy'] 'splitter': ['best','random']

Table 3.7 - Hyperparameters explored for each model

As we have two target variables *Anxiety_Level* and *Depression_Level,* so we explored hyperparameters using GridSearchCV for predicting both the target classes.

3.10.1 Best hyperparameters for predicting Anxiety Level

The following Table-3.8 describes the best hyperparameters tuned for each model, to predict *Anxiety Level*, along with the best score.

Model	Best Parameters	Best Score
Support Vector Classifier (SVC)	{'C': 1, 'kernel': 'rbf'}	0.427
Random Forest Classifier	{'criterion': 'gini', 'n_estimators': 100, 'random_state': 3}	0.416
Logistic Regression	{'C': 1, 'random_state': 3}	0.405
K-Neighbors Classifier (KNN)	{'algorithm': 'auto', 'n_neighbors': 20, 'weights': 'distance'}	0.411
LightGBM (LGBM)	{'max_depth': -1, 'random_state': 5}	0.356
Decision Tree	{'criterion': 'gini', 'splitter': 'best'}	0.361

Table 3.8 - Best Hyperparameters for predicting *Anxiety Level*

3.10.2 Best hyperparameters for predicting Depression Level

The following Table-3.9 describes the best hyperparameters tuned for each model, to predict *Depression Level*, along with the best score.

Model	Best Parameters	Best Score
Support Vector Classifier (SVC)	{'C': 1, 'kernel': 'rbf'}	0.377
Random Forest Classifier	{{'criterion': 'entropy', 'n_estimators': 50, 'random_state': 6}	0.322
Logistic Regression	{'C': 1, 'random_state': 3}	0.333
K-Neighbors Classifier (KNN)	{'algorithm': 'auto', 'n_neighbors': 15, 'weights': 'distance'}	0.344
LightGBM (LGBM)	{'max_depth': -1, 'random_state': 5}	0.295
Decision Tree	{'criterion': 'gini', 'splitter': 'best'}	0.268

Table 3.9 - Best Hyperparameters for predicting *Depression Level*

Chapter 4

Results

We used six different classifiers and the dataset was trained on the aforementioned classifiers. To determine the best classifier, they were assessed against a variety of performance indicators. We evaluated the classifiers performance using four distinct metrics. They are: Accuracy, Precision, Recall, F1-Score.

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

F1 Score is the weighted average of Precision and Recall.

$$F1\ Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4.4}$$

To measure the overall performance of each of the six classifiers, five-fold cross validation (cv=5) is used. The data set is divided into five folds in this case. The first iteration uses the first fold to test the model, while the remaining folds are utilized to train the model. The second iteration uses the second fold as the testing set and the remaining folds as the training set. This process is repeated until each fold of the five folds has been used as the testing set. This is an effective cross validation approach to measure the performance of a model.

Performance of the classifiers in predicting *Anxiety Level* based on the best hyperparameters is summarized in Table-4.1.

Model	Accuracy	Precision	Recall	F1 Score
Support Vector Classifier (SVC)	0.51	0.53	0.77	0.62
Random Forest Classifier	0.35	0.44	0.55	0.48
Logistic Regression	0.41	0.70	0.54	0.61
K-Neighbors Classifier (KNN)	0.51	0.62	0.77	0.67
LightGBM (LGBM)	0.41	0.50	0.73	0.59
Decision Tree	0.14	0.23	0.27	0.25

Table 4.1 - Performance of the Classifiers in predicting *Anxiety Level*

On the other hand, performance of the classifiers in predicting the other target variable, *Depression Level* based on the best hyperparameters is summarized in Table-4.2.

Model	Accuracy	Precision	Recall	F1 Score
Support Vector Classifier (SVC)	0.32	0.56	0.69	0.62
Random Forest Classifier	0.35	0.67	0.62	0.57
Logistic Regression	0.30	0.50	0.54	0.52
K-Neighbors Classifier (KNN)	0.51	0.73	0.73	0.73
LightGBM (LGBM)	0.35	0.67	0.54	0.56
Decision Tree	0.22	0.50	0.38	0.43

Table 4.2 - Performance of the Classifiers in predicting *Depression Level*

Chapter 5

Conclusion

In this work, we have surveyed school, college and university students from 30 different institutes of Bangladesh to analyze their mental health (anxiety and depression) after the reopening of educational institutions. Each participant had to fill up a mental health questionnaire that contained 52 personalized questions, including 7 questions from GAD-7 scale to measure anxiety level, and 9 questions from PHQ-9 scale to measure depression level. The survey was conducted through online and offline methods and a total of 183 instances of mental health data have been collected over a period of 1.5 months. The aim of this work is to predict anxiety severity and depression severity amongst students after the reopening of schools based on 34 set of featured questions.

Several preprocessing steps have been performed to make the dataset ready for the machine learning models to be run. Six classifiers have been used in our work – Support Vector Classifier (SVC), Light Gradient Boosting Machine (LGBM), Random Forest, K-Nearest Neighbor (KNN), Logistic Regression and Decision Tree, to predict the two classes *Anxiety Level* and *Depression Level.* Hyperparameter Tuning have been performed individually, for the two target variables, to identify the best parameters of each classifier. After applying the classifiers, K-Nearest Neighbor (KNN) gave the best accuracy of 51% for predicting *Anxiety Level* and 51% for predicting *Depression Level.* For the prediction of *Anxiety Level,* support vector classifier (SVC) was the 2nd best classifier model and gave an accuracy of also 51%. On the other hand, for the prediction of *Depression Level,* Random Forest Classifier and LGBM both gave accuracy of 35% as the 2nd best classifier.

Our accuracy and other metrics are low because of lack of data in our dataset. We tried to augment data by producing synthetic tabular data using Generative Adversarial Network (GAN) and increased our dataset to 1900 instances including synthetic data, but the classifiers had even worse results on the synthetic dataset. For further work, we would like to collect more data of our survey to ensure that the machine learning classifiers being used have enough data to effectively learn from the dataset and correctly predict the target variables.

References

[1] Unicef, "Education disrupted: The second year of the COVID-19 pandemic and school closures," Unicef, September 2021. [Online]. Available: https://data.unicef.org/resources/education-disrupted/. [Accessed Septenber 2022].

[2] M. M. Jasim, "School closure longest in Bangladesh, learning vacuum alarming: Educationists," The Business Standard (TBS), 18 February 2022. [Online]. Available: https://www.tbsnews.net/bangladesh/education/school-closure-longest-bangladesh-learning-vacuum-alarming-educationists-372592. [Accessed August 2022].

[3] R. Spitzer, K. Kroenke, J. Williams and L. B., "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7," *Arch Intern Med.,* vol. 166, no. 10, p. 1092–1097, 2006.

[4] K. K, R. L. Spitzer and J. B. Williams, "The PHQ-9: validity of a brief depression severity measure," *J Gen Intern Med,* vol. 16, no. 9, pp. 606-13, 2001.

[5] W. Jingyi, W. Yingying, L. Haijiang, C. Xiaoxiao, W. Hao, H. Liang, X. Guo and C. Fu, "Mental Health Problems Among School-Aged Children After School Reopening: A Cross-Sectional Study During the COVID-19 Post-pandemic in East China," *Frontiers in Psychology,* vol. 12, no. 773134, 2021.

[6] R. Ziyuan, X. Yaodong, G. Junpeng, Z. Zheng, L. Dexiang, C. M. H. Roger and S. H. H. Cyrus, "Psychological Impact of COVID-19 on College Students After School Reopening: A Cross-Sectional Study Based on Machine Learning," *Frontiers in Psychology,* vol. 641806, p. 2, 29 April 2021.

[7] C. Zhou, R. Li, M. Yang, S. Duan and C. Yang, "Psychological Status of High School Students 1 Year After the COVID-19 Emergency," *Frontiers in psychiatry,,* 2021.

[8] E. Mostafiz, M. T. Husein, A. A. Mahfug and A. B. M. Himal, *52 set questions of our Mental Health Survey Questionnaire,* Dhaka, 2022. Available online: https://drive.google.com/file/d/1kExEvx7NUyp3WwqTaX-v1wz8N1uy_Nbg/view?usp=sharing

[9] M. Erfan, M. Abdullah Al, H. Md Toufique and A. B. M. Himel, "CSE498R Mental Health Survey Dataset," 2022. Available online: https://docs.google.com/spreadsheets/d/1780tvSRPZf4RiS4WL8QCRALp5QtX0-IUUDoqesljYDM/edit?usp=sharing

[10] H. Md Toufique and M. Erfan, "Our .ipynb notebook of detecting depression and anxiety among Bangladeshi students," 2022. Available online:

https://colab.research.google.com/drive/1Q54Df93qL6w3tgEcE7TlqLdAmhkB6HM-?usp=sharing

[11] J. Brownlee, "Ordinal and One-Hot Encodings for Categorical Data," Machine Learning Mastery, 17 August 2020. [Online]. Available: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/. [Accessed August 2022].

[12] G. L. Team, "Label Encoding in Python Explained," My Great Learning, 16 December 2021. [Online]. Available: https://www.mygreatlearning.com/blog/label-encoding-in-python/. [Accessed August 2022].

[13] L. scikit-learn, "sklearn-labelEncoder," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html?highlight=labelencoder.

[14] S. L. Library, "Scikit Learn Library," [Online]. Available: https://scikit-learn.org/stable/index.html.

[15] MinMaxScaler. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[16] TrainTestSplit, "sklearn.model_selection.train_test_split," sk-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html?highlight=train_test_split#sklearn.model_selection.train_test_split.

[17] Heavy.AI, "Feature Selection Definition," HEAVY.AI, [Online]. Available: https://www.heavy.ai/technical-glossary/feature-selection.

[18] J. Benesty, J. Chen, Y. Huang and I. Cohen, "Pearson correlation coefficient.," in *Noise reduction in speech processing*, Berlin, 2009.

[19] M. Milecia, "SVM Machine Learning Tutorial – What is the Support Vector Machine Algorithm, Explained with Code Examples," FreeCodeCamp, 1 July 2020. [Online]. Available: https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/. [Accessed September 2022].

[20] L. Andy and W. Matthew, "Classification and Regression by RandomForest," *Forest,* vol. 23, 2001.

[21] J. Tolles and W. Meurer, "Logistic Regression: Relating Patient Characteristics to Outcomes," *JAMA The Journal of the American Medical Association,* vol. 316, no. 5, p. 533, 2016.

[22] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory,* vol. 13, no. 1, pp. 21-27, 1967.

[23] G. Ke, Q. Meng, T. Finely, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems 30,* vol. NIPS 2017, p. 3149–3157, 2017.

[24] J. Jordan, "Hyperparameter tuning for machine learning models.," Jeremy Jordan, 2 November 2017. [Online]. Available: https://www.jeremyjordan.me/hyperparameter-tuning/. [Accessed September 2022].