

Predictive Analysis on Depression among University Students in Bangladesh

By

Syed Aref Ahmed

19201124

Khondokar Jamal E Mustafa

19241008

Ibtesum Arif

19201054

MD. Nafis Tahmid

19301053

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
Brac University
September 2023

© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/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. I/We have acknowledged all main sources of help.

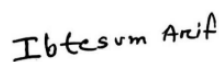
Student's Full Name & Signature:



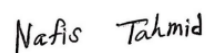
Syed Aref Ahmed
19201124



Khondokar Jamal E Mustafa
19241008



Ibtesum Arif
19201054



MD. Nafis Tahmid
19301053

Approval

The pre-thesis 1 report titled “Predictive Analysis on Depression among University Students in Bangladesh” submitted by

1. Syed Aref Ahmed (19201124)
2. Khondokar Jamal E Mustafa (19241008)
3. Ibtesum Arif (19201054)
4. MD. Nafis Tahmid (19301053)

Of Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of BSc. In Computer Science and Engineering on [Date-of-Defense].

Examining Committee:

Supervisor:
(Member)



Md. Shahriar Rahman Rana
Lecturer, Dept. of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)



Rafeed Rahman
Lecturer, Dept. of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam
Associate Professor, Dept. of Computer Science and
Engineering
Brac University

Departmental Head:
(Chair)

Ms. Sadia Hamid Kazi
Chairperson, Lecturer, Dept. of CSE
Brac University

Abstract

The identification of depression is done by medical practitioners based on mental status questionnaires and the patient's self-reporting. Apart from the methods being highly dependent on the patient's current mood, people who go through mental disorders seek mental help reluctantly. Universities always promise scholars a promising career in their domains. However, the academic competition, peer pressure, isolation and many other factors could put a student in a state of depression. In this research, we propose a big data analytics template to detect depression among university students. Asserting again, since isolation and separation are believed to have the most dramatic effect on the pupils, the framework also models the correlation between these factors and depression. To conclude, the journal evaluates the performance of the proposed framework on a massive real dataset collected from different university students of Bangladesh and proves that the accuracy of the machine learning models outperforms traditional techniques for detecting depression in universities.

Keywords: depression; mental health; academic competition; peer pressure; isolation; big data; university students of Bangladesh; accuracy; detection;

Dedication

In the past, there was a lot of stigma associated with depression and other mental health disorders. Many people were afraid to seek treatment because they thought it would make them look weak or weak-minded. This stigma still exists today, especially among those who suffer from depression.

The purpose of this research paper is to demonstrate how the stigma surrounding depression can be reduced by educating others about this disease. It is hoped that by understanding the causes and effects of depression, we can encourage more people to seek help when they are suffering from this condition.

Acknowledgement

This is an optional page. Use your choice of paragraph style for text on this page (1_Para shown here).

Table of Contents

Declaration.....	ii
Approval	iii
Abstract.....	iv
Dedication	v
Acknowledgement	vi
Chapter 1: Introduction	1
1.1 What is Depression?	1
1.1.1 <i>Effects of Depression Among University Students</i>	1
1.1.2 <i>Use of Predictive Analysis to Detect Depression</i>	2
1.2 Problem Statement.....	2
1.3 Research Objective	3
Chapter 2: Literature Review.....	4
2.1 Paper Review	4
2.2 Major Findings from Review	10
Chapter 3: Methodology.....	11
3.1 Input Data.....	12
3.2 Data Pre-processing	12
3.3 Feature Scaling.....	15
3.4 Implementation and Results	16
3.5 Data Pre-processing	22
Conclusion	25
References.....	26

Chapter 1: Introduction

1.1 What is Depression?

Depression is a serious illness that affects your ability to function normally. You may not feel like yourself, and you might feel sad or anxious most of the time. Depression can be mild or severe, and it can last for months or years.

Depression is a medical condition that has many different causes. It can develop for no apparent reason and it often occurs in people who have other health problems. Depression is common in women during pregnancy, with about one in 10 women experiencing depression during their childbearing years, compared to about one in 20 men. Depression is also common among older adults (especially those over age 65). In the 21st Century depression is getting common among university students due to various factors.

Depression affects your thoughts, feelings and behaviors. It may also affect your relationships with others, especially with family members and friends who are close to you. Your symptoms may cause serious problems in your social life and work performance because they make it hard for you to function normally at home or at work. Depression can interfere with everyday activities such as getting ready for work or going out to eat lunch with friends on the weekends.

1.1.1 Effects of Depression Among University Students

Depression is a mental illness that can bring great distress, suffering and hopelessness to people. It is common in all age groups, but affects young people more than any other age group.

University Students who are depressed may not be able to concentrate on their studies. They may even start missing classes. This can cause problems in the future, as they may not be able to keep up with their classmates and will have difficulty getting good grades.

Some students become so overwhelmed by their depression that they stop eating or sleeping properly. They may also drink too much alcohol or take drugs in order to cope with their feelings of hopelessness and despair. Students who are depressed often feel alone and isolated from others, which makes it even harder for them to seek help from family members, friends or teachers.

1.1.2 Use of Predictive Analysis to Detect Depression

The term "predictive analysis" is used to describe the process of identifying relationships between variables or events. In this case, it will be used to identify relationships between depression and other factors such as age, gender, marital status, employment status and so on.

This is a very useful technique that can help us understand how certain people are more prone to depression than others. This can be used by psychologists in order to design programs that they want to implement in their own practice. It can also help them identify the most effective treatment strategies for specific individuals based on their unique characteristics.

1.2 Problem Statement

Currently, depression is a very big issue university students are dealing with in their day-to-day life. People's lives have gotten much faster over the years and lot of things can affect a university to lead into depression. We chose our topic to detect a pattern of why university students are getting into depression so that those who need help can be helped.

Our main focus with this research is to identify a depressed person based on the various factors from the dataset we are currently working on. Our plan is to apply different machine learning models to find out which model will have the highest accuracy for our research. As a result,

we are going to use the most effective machine learning model and detect patterns through predictive analysis to detect depressed university students in Bangladesh.

1.3 Research Objective

Depression is a crippling mental health problem that is still not appropriately treated in many countries, including Bangladesh. People's hesitation to seek expert assistance is one of the main causes of this. This hesitation results from a lack of knowledge about the illness and its consequences. Unfortunately, a growing number of young people, particularly university students in Bangladesh, are succumbing to this pernicious illness as a result of this lack of understanding of depression. The figures show rising suicide rates, and depression is now recognized as a major underlying cause.

The causes of this alarming trend are numerous. A factor in preventing people from expressing their troubles and getting help is the social stigma associated with mental health concerns. In addition, the general social view of mental health frequently supports the notion that depression is a condition that can be "snapped out of" or is not a true illness. The results of this growing depression epidemic are disastrous. Students are particularly vulnerable since they are juggling the difficulties of higher education and maturity. Their challenges with mental health may be made worse by the weight of societal expectations, scholastic pressure, and the ambiguity of the future.

A person's well-being depends both on their physical and mental health. So, it is very important in our society to address the situation of depression. The growing rate of depression among university students in Bangladesh is very alarming and needs to be tackled immediately. Our research focuses on improving the identification and assisting students who are at risk of depression. The goals of this research are:

1. Provide insights to improve student welfare sector in universities in Bangladesh

2. Develop a model which can detect depression
3. Enhance focus on mental health of university students
4. Analyze patterns of various factors that has a high chance to lead university students to depression

Chapter 2: Literature Review

One of the best tools we have right now to fight depression is a wonder of computer science and a branch of artificial intelligence, machine learning. Machine learning is great at dealing with complex datasets and figuring out how to model these different islands. In machine learning, we start with a goal in mind. We give the machine all the data, any data that we can get. After putting all the data together in a machine, the machine attempts to learn as best as it could. At the same time, a model that takes some of these variables all together, will learn what someone or something responds to that particular goal. Moreover, several techniques have been implemented to detect depression. From them, some machine learning approaches on numerous datasets are visible.

2.1 Paper Review

Helbich et al. deduced depression severity through the depression module of the Patient Health Questionnaire (PHQ-9). They improved the accuracy of the classifier by considering some attributes such as physical neighborhood environment and social neighborhood environment. Furthermore, to deduce the severity of depression and to generate the targeted PHQ-9 score artificial neural network, random forest classifier and gradient boosting machine were used. The models generated the PHQ-9 scores based on several factors such as age, gender, marital status, employment status and a few more. The results showed the higher the PHQ-9 scores were, the more depressed each individual was. On the contrary, a lower PHQ-9 indicated a happier human being. Furthermore, the study showed that the employed and the person with a

higher education or income had a significantly lower PHQ-9 score than the people who had worse economic conditions. The authors concluded by saying that social environment and individual attributes were most important factors as opposed to physical environment. [1]

A paper titled “Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States” by Kumar and Chong also conducted similar study [2] , that considers attributes from weather and physiological factors in the data set. After running feature selection Temperature, Atmospheric, Atmospheric and Season were found to be four most important weather attributes. In addition to that, Tana et al. performed a statistical analysis of prevalence of depression through infodemiology [3] . Prevalence of depression related terms and keywords on twitter was plotted on yearly, season and daily basis. It was found that depression-related terms were most tweeted during cold seasons.

In another study Shahriar et al. aimed to detect clinical depression by considering some common attributes and by implementing some machine learning algorithms. Shahriar figured out random forest generated the highest accuracy of 83.80% while K-means clustering gave the lowest output of 68.15%. However, after the models were applied on a primary dataset, he was able to create another intermediate dataset based on the accuracy of the algorithms. In this case, the improvement in precision, performance metrics and the machine learning algorithms were promising. For instance, the accuracy of KNN classifiers increased to 93% from 83.52% and so on. After that, he was able to do some statistical analysis on some factors like education level, marital status, gender, assets and so forth. from the improved data. To sum up, Shahriar tried to deduce clinical depression from the analysis of socio-economic attributes. [4]

Apart from socio-economic factors, the newly global-spread virus COVID-19 also contributed to the deterioration of mental health by a marginal extent among people during lockdown. A

research was done before nation-wide lockdown due to Corona Virus in Liaoning Province, China by Zhang and Ma [5] and they concluded that COVID-19 pandemic was associated with mild stressful impact in their sample.

HosseiniFard et al. [6] classified patients suffering from depression and normal subjects by using machine learning algorithms and nonlinear features from EEG signals. The research work had been set up with two class of subjects where 45 were depressed subjects and other 45 were normal subjects from the interested participants. K-nearest neighbor, linear discriminant analysis and logistic regression were used to separate the identified groups. Interestingly, the accuracy of logistic regression improved from 83.3% to 90% by combining all other nonlinear features and by reapplying those into classifiers. However, to get a proper insight of the EEG signal of depressed subjects, some non-additional attributes were also needed. EEG signals vary significantly from a normal person to a mentally disturbed person, it could find depressed people quite accurately. Stressing again, EEG signals are generated in different parts of the brain due to depression. Therefore, by investigating those targeted regions of the brain, curing depression would take a very little time. To sum up, this paper could be a significant tool to detect depression by applying nonlinear analysis on EEG signals.

Razavi et al. [7] detected depression by using continuous cell phone usage patterns. Moreover, they improved the accuracy of the classifier by taking into consideration some explicit attributes such as age, gender and so on. As a result, the authors propose to use mobile data, mental health apps and user internet history to monitor ongoing treatments. They found out that a random forest classifier generated the highest accuracy of 81% in detecting depression compared to other machine learning algorithms.

Although our target audience is university students in Bangladesh, studies related to other groups of people will help us understand better. Sau and Bhakta [8] studied predicting anxiety and depression among elderly people. This was one of the studies, where a relatively large amount of data was used. Also, due to large data sets, it was possible to train the data set by various Machine Learning algorithms. Like most other cases, the Random Forest algorithm performed better than the remaining others.

Similar to the previous study by Sau and Bhakta, Haque et al. [9] conducted their study on predicting mental health conditions of children aged between 4 - 17 years of age in binary format (depressed or not depressed). The topic of mental health conditions is a very rarely discussed topic. Although the sample size was also relatively large, the data set was heavily imbalanced. So, accuracy was not a good measurement to measure performance, AUC scores and ROC(AUC) scores were also measured. Random Forest algorithm worked best here also.

In order to observe how imbalanced data sets, affect Machine Learning models, Jeni et al. [10] three major datasets were used (Cohn-Kanade, RU-FACS, and McMaster Pain Archive). It found that with exception of area under the ROC curve, all performance metrics were attenuated by imbalanced distribution.

The goal of the study by Schwanz et al. [11] titled "Self-Reliance and Relations with Parents as Predictors of Anxiety and Depression in College Students" was to predict anxiety and depression in college students using these factors. The participants were 153 students. Each of the 153 individuals received a score for Self-Reliance T on the BASC, Parent Relations T on the BASC, Trait Anxiety T on the STAI, and Depression T on the HDI. Measures' descriptive statistics and correlation matrix. Participants' levels of self-reliance and self-reported perceptions of their relationships with their parents were assessed using the Behavior Assessment System for Children 2nd Edition Self Report of Personality College form (BASC-

2 SRP-COL). A tool used to identify and assess particular symptoms of Major Depression as defined by DSM-IV is the Hamilton Depression Inventory (HDI). The whole form is 23 items long. Also used was STAI. Self-reliance scores contributed more to the prediction of each of the outcome variables than parent relationships, according to multiple regression analyses that showed them to be significant predictors of anxiety and depression. Similar to this paper Rois et al. [12] conducted a predictive analysis paper to predict Prevalence and predicting factors of perceived stress among university students in Bangladesh. The six prevailing factors were: pulse rate, SBP, DBP, sleep status, smoking, background [department]. Among multiple training algorithms, Random Forest performed the best. There is another paper published by Choudhury et al. [13] that works on predicting depression among Bangladeshi students, but the survey had only a limited number of questions.

A paper published by Kumar and Chong [14] takes a completely different approach in finding effective data for correlation analysis. The paper identified depression by analyzing whole-brain functional connectivity. Support vector classification was used to find a solution to the problem. Furthermore, dataset was created by analyzing different parts of the brain such as cerebellum, affective network, visual cortical areas and so on from volunteers. Meanwhile, SVM predicted depression with an accuracy of 94.3% in this case. Asserting again, their report generated the highest accuracy among all the papers on this subject. Moreover, the probability of the accuracy being wrong was only < 0.0001 in this case. Moreover, data of the functional regions were also needed for the research. Asserting again, the supporting vector machine gave an accuracy of 92.5% in this case.

“Developing Depression Symptoms Prediction Models to Improve Depression Care” by Jin et al. [15] significantly improved the rate of true positive for predicting major/severe depression by using multinomial logistic regression model used variables that are highly correlated with outcome but with low inter-correlation as predictors.

Although there have been many studies on predicting depression and stress, there are many more to explore related to detecting depression among people. One such study is done by Taawab et al. [16], that detects self-esteem level and depressive indication. Due to different parenting styles using supervised learning techniques. The researchers created their own dataset from about 500 survey responses. Apart from using various python libraries, supervised models such as Logistic Regression, Gradient Boost Classifier and Bi-Directional LSTM provided better accuracy than other techniques. With 80% training data and 20% testing data, the Gradient Boosting Classifier (GBC) gave the best performance with an accuracy of 95% and F1 score of 83.28%. Furthermore, word embedding techniques also generated a performance with 83.21% accuracy and 83.50% recall score in successfully detecting depression and self-esteem problems among the youths.

Similar to that, Raze et al. [17] conducted study on another less researched topic, early detection of postpartum depression in Bangladesh. This paper measures postpartum depression level of mothers in Bangladesh. (0 - 7: Low, 8-15: Medium, 16-23: High). Data set was obtained via a survey of 150 mothers. Synthetic minority oversampling approach (SMOTE) was used to correct the data imbalance. The data set was tested via various ML models and each model's performance was measured via various performance metrics. Random forest performed the best.

In contrast to these studies, our approach will be a bit different. We look forward to developing different kinds of machine learning models to detect depression of a university student by taking some socio-economic data. At the same time, we will also create another immediate dataset to enhance the performance of the models.

2.2 Major Findings from Review

After reading various research papers similar to our topic, we have gathered some major findings which are given below.

- i. There are various machine learning algorithms that can be used in training models such as Random Forest, Multinomial Logistic Regression, Logit Boost, Random Forest, Support Vector Machines, KNN, Decision Tree, Naive Bayes, SVM, Bagging, Extra Trees, Ada Boosting, Gradient Boosting, TPOT Classifier, XGM, DT, Gaussian NB. Out of all these, Random Forest performs most efficiently.
- ii. Some algorithms may need feature scaling or selection. Pearson's correlation and Boruta algorithms are mainly used for feature scaling or selection.
- iii. Performance is measured with accuracy in most cases. However, in imbalanced data sets ROC curve gives more accurate results.
- iv. There have been many studies done with detecting depression among people. But, the topic of depression and mental health is vast, so there is more to study in this field.
- v. Machine learning models work better when the data set is large. It can be trained with various algorithms. Also, in almost all studies, the data have been collected via authors themselves. So, when working with depression, it's better to make data sets by the authors themselves.

Chapter 3: Methodology

The aim of our research is to enhance student welfare and create a model for identifying depression by exploring the factors behind its cause. In Bangladesh, depression is a significant problem worsened by social stigma and a lack of awareness. University students impacted by it particularly.

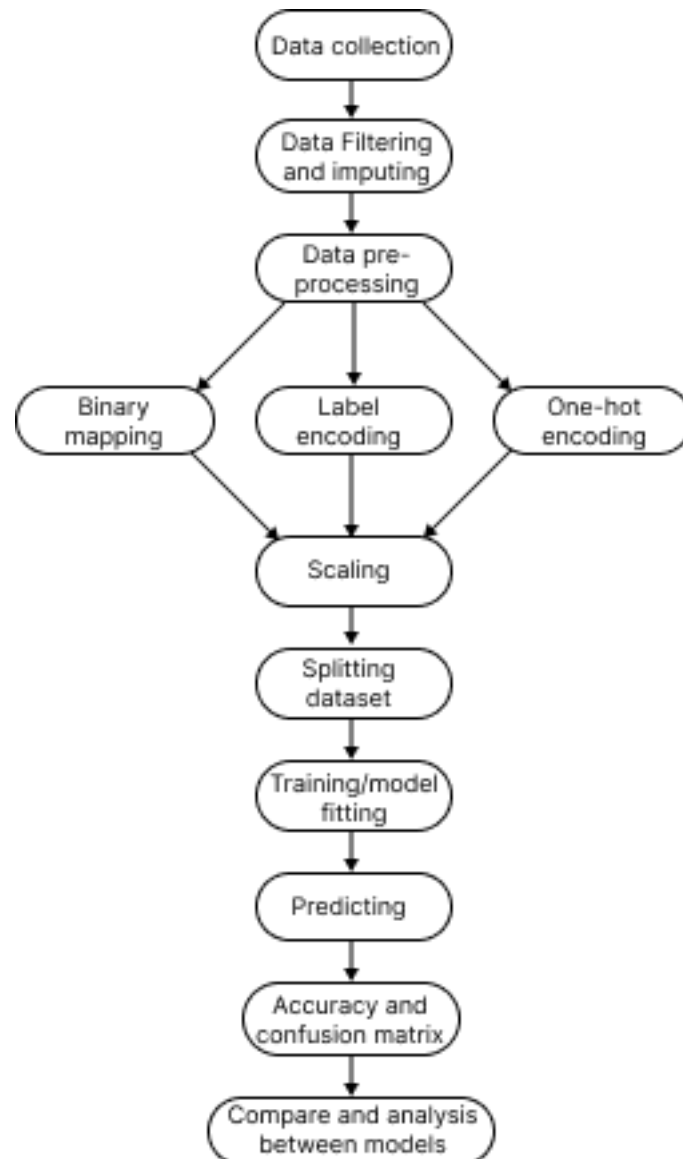


Figure 1: Flowchart of the proposed detection model.

3.1 Input Data

Datasets in Machine Learning are important for two reasons mainly: to train machine learning models and to provide a benchmark for measuring the accuracy of those trained models. Creating the dataset for our predictive analysis model was a big challenge. As we desire to analyze the trends and patterns of mental health of students of current time, we decided to create our own dataset by surveying students from various universities via google form. We needed to extract as much as relevant information as possible but we also needed to maintain privacy and comfort of those participant. Therefore, the form was made anonymous and we asked 26 queries to each participant. A total of 612 responses were obtained.

3.2 Data Pre-processing

As usual, all the entries in the dataset are string type. Among the 26 queries (each query corresponds to a column), the final query asks if the student is suffering from any depression or not. This query is considered as target class for our dataset and it is a binary class categorical query as the model will perform binary classification on the dataset. Among the remaining 25 queries (features of the dataset) 4 queries had numeric options, 8 queries were multi-class categorical queries and remaining 13 were binary class categorical queries. The table below illustrates the data type of each feature.

Binary	Multiclass	Numeric
Gender	University year	Age
Born in capital	Educational background	Number of family members
Earning as of now	Relationship status	Number of children
Satisfied with academy	Monthly income	Sleeping hours
Any physical disabilities	Monthly living expense	
Ever been in a road accident	Social gatherings a week	
Any childhood trauma	Time on social media	
Taking any medication	Social life satisfaction	
Religious person		
Any indoor/fun activity		
Sports/gym		
Coffee person		
Any addictive substance		

Table 1: Queries of our survey, grouped by data types

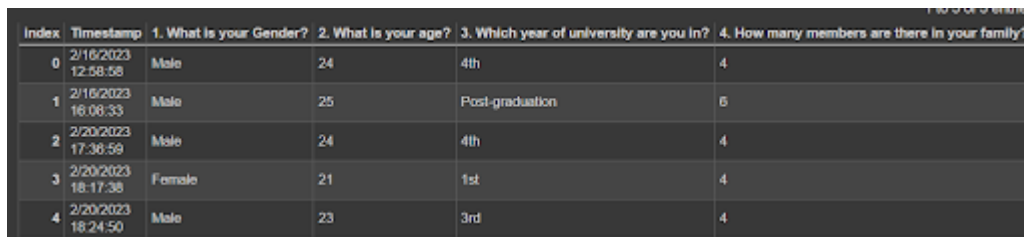
Before data preprocessing the initial feature names, which were the questions of the form, were changed to more appropriate, variable-like feature names. Also, we checked for missing values in the data set and we found that only one column had 61 missing values (number of children). We imputed missing values by taking mode of that feature.

The binary categorical features had two options: yes and no (the queries were agree/disagree type). We replaced the values with 0 and 1 by using the replace method of the panda package.

All the numeric values were discrete integer values with finite range. The values of age feature and sleeping duration feature were compressed to smaller groups. Next, these two features, the remaining integer valued features (family members and number of children) and the university year column (multiclass feature) were one-hot encoded, as there was no specific order of significance in the values.

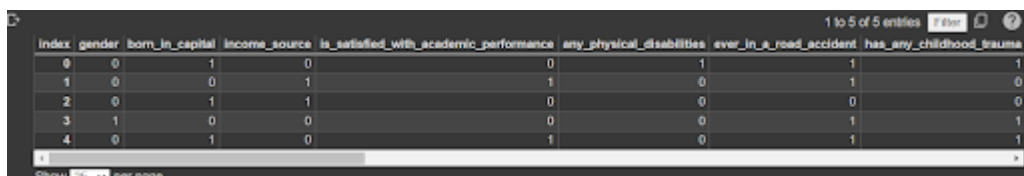
Lastly, the remaining 7 multi-class categorical features were label encoded, as there was an order of rank among the values of those features.

After performing all the preprocessing steps our dataset gets drastically changed, where each contains either binary values or discrete numeric values. This transformed dataset is now prepared for feature scaling and model implementation.



Index	Timestamp	1. What is your Gender?	2. What is your age?	3. Which year of university are you in?	4. How many members are there in your family?
0	2/16/2023 12:58:58	Male	24	4th	4
1	2/16/2023 16:06:33	Male	25	Post-graduation	6
2	2/20/2023 17:36:59	Male	24	4th	4
3	2/20/2023 18:17:38	Female	21	1st	4
4	2/20/2023 18:24:50	Male	23	3rd	4

Figure 2: Dataset before pre-processing



Index	gender	born_in_capital	income_source	is_satisfied_with_academic_performance	any_physical_disabilities	ever_in_a_road_accident	has_any_childhood_trauma
0	0	1	0	0	1	1	1
1	0	0	1	1	0	1	0
2	0	1	1	0	0	0	0
3	1	0	0	0	0	1	1
4	0	1	0	1	0	1	1

Figure 3: Dataset after pre-processing

3.3 Feature Scaling

Feature scaling is a technique to standardize or normalize the features into a smaller scale or range. The goal of feature scaling is to bring all features to a common scale without distorting the relative differences between their values. Although feature scaling is not important for tree based and decision-based algorithms, algorithms such as Logistic Regression, Support Vector Machine and K-Nearest Neighbors greatly benefit from standardization feature scaling.

Standardization, also known as Z-score normalization, scales the features so that they have a "mean (average) of 0" and a "standard deviation of 1". "Mean of 0" means the average of all values of a feature is equal to 0 or very close to 0. "Standard deviation of 1" that the spread of the values in the standardized dataset is similar to what you'd expect from a standard normal distribution (a bell-shaped curve with a mean of 0 and a standard deviation of 1).

Standardization scales the values in such a way that the spread becomes consistent, and the standard deviation is 1.

We performed feature scaling in a training and testing set via StandardScaler class imported from sklearn.preprocessing package. The formula for standardization is:

$$x(\text{stand}) = \frac{x - \text{Mean}(x)}{\text{Standard Deviation}(x)}$$

- $x(\text{stand})$ = the standardized value of the feature.
- x = the original value of the feature
- $\text{Mean}(x)$ = the mean (average) of the feature (calculated by the StandardScaler).
- $\text{Standard deviation}(x)$ = the standard deviation of the feature (calculated by the StandardScaler).

3.4 Implementation and Results

From the literature review we concluded that there are few classifications algorithms that have much higher accuracy than most other algorithms. Those algorithms are:

- Logistic Regression
- Decision tree
- Random Forest
- AdaBoost
- Gradient Boost
- Support Vector Machine
- KNN
- Naive Bayes

In our research paper we will implement all these 8 algorithms and measure their performances on various metrics such as accuracy score, confusion matrix etc.

Accuracy score is a very simple, yet very important concept for determining effectiveness of any machine learning model. Accuracy score is simply the ratio of number of right predictions and total number of samples.

$$\text{Accuracy score} = \frac{\text{Number of correct prediction}}{\text{Sample size (n)}}$$

Confusion matrix is a two-dimensional matrix of size 2x2 that illustrates the correctness of predictions.

	[Predicted negative]	[Predicted positive]
[Actual negative]	True Negative (TN) - Predicted false, actual was false	False Positive (FP) - Predicted true, actual was false
[Actual positive]	False Negative (FN) - Predicted false, actual was true	True Positive (TP) - Predicted true, actual was true

Table 2: Confusion matrix format

Now, we will analyze the 8 mentioned algorithms.

- Logistic Regression (LR): This model was imported from `sklearn.linear_model`.

Dataset was divided into training and testing sets in a 4:1 ratio. Then the model was fitted into the training set and predicted values were stored. The model worked with 99.2% accuracy. The confusion matrix was:



Figure 4: Confusion matrix for LR

- Decision Tree (DT): The procedure was very similar to the previous model implementation, except `DecisionTreeClassifier` was imported from `sklearn.tree` package. The accuracy was 97.6%. The confusion matrix that was obtained:

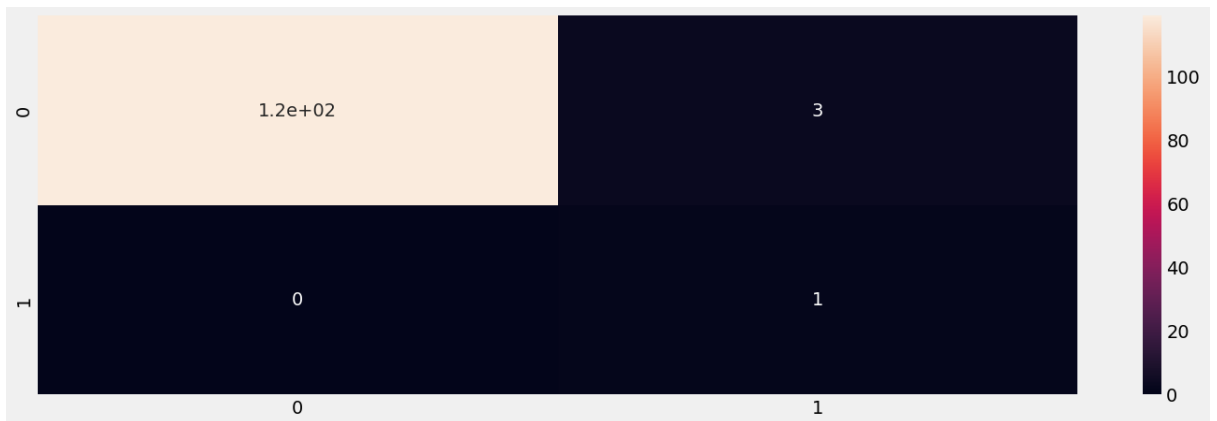


Figure 5: Confusion matrix for DT

Also, the tree that was generated:

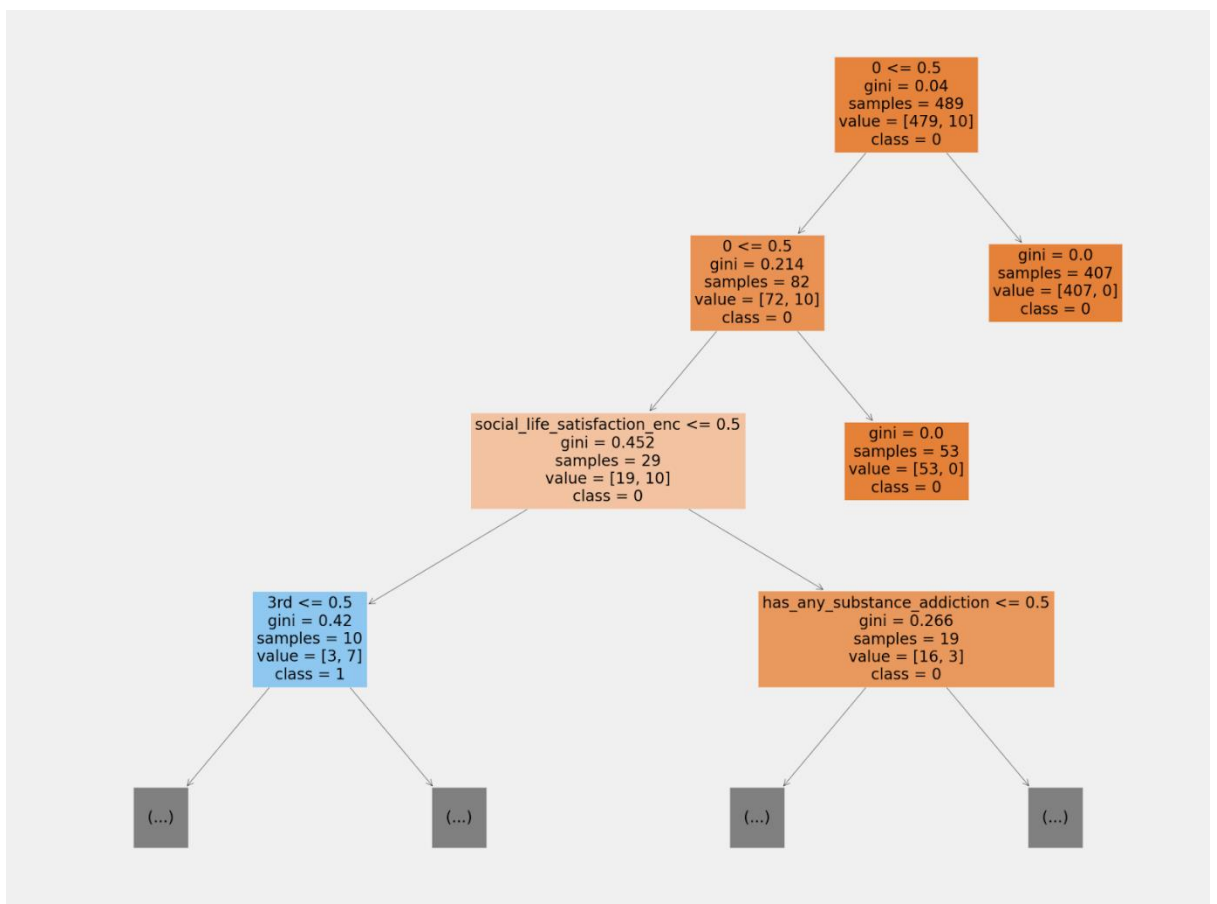


Figure 6: Tree created by DT classifier

- Random Forest (RF): RandomForestClassifier was imported from sklearn.ensemble package. The accuracy was 100%. The confusion matrix:

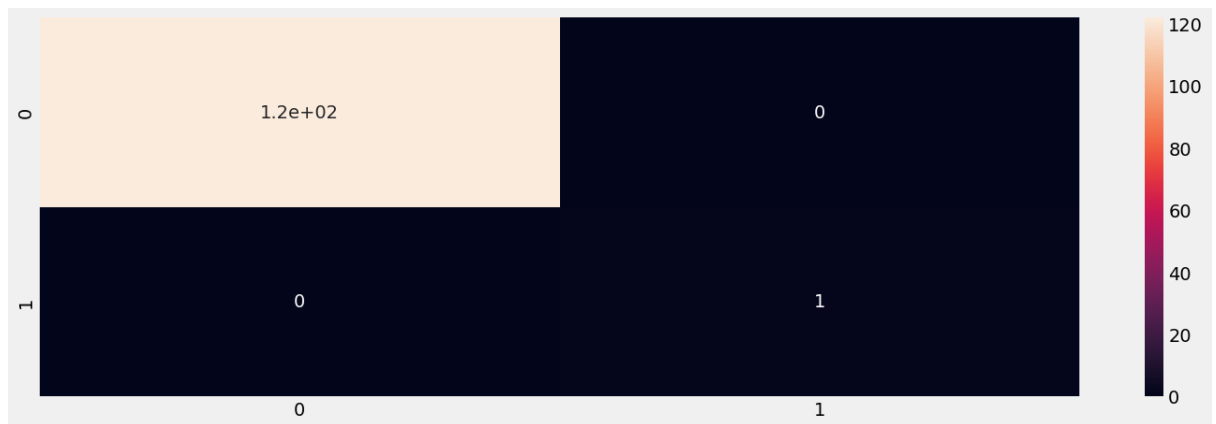


Figure 7: Confusion matrix for RF

- AdaBoost (Ada): AdaBoostClassifier from sklearn.ensemble. Accuracy: 100%.

Confusion matrix:

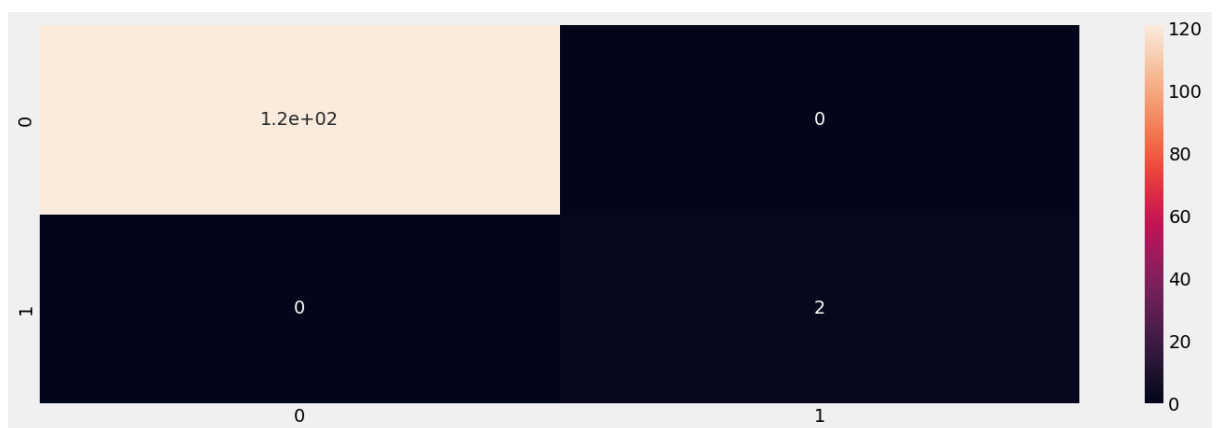


Figure 8: Confusion matrix for ADA

- Gradient Boost (GB): GradientBoostingClassifier from sklearn.ensemble. Accuracy: 98.4%. Confusion matrix:

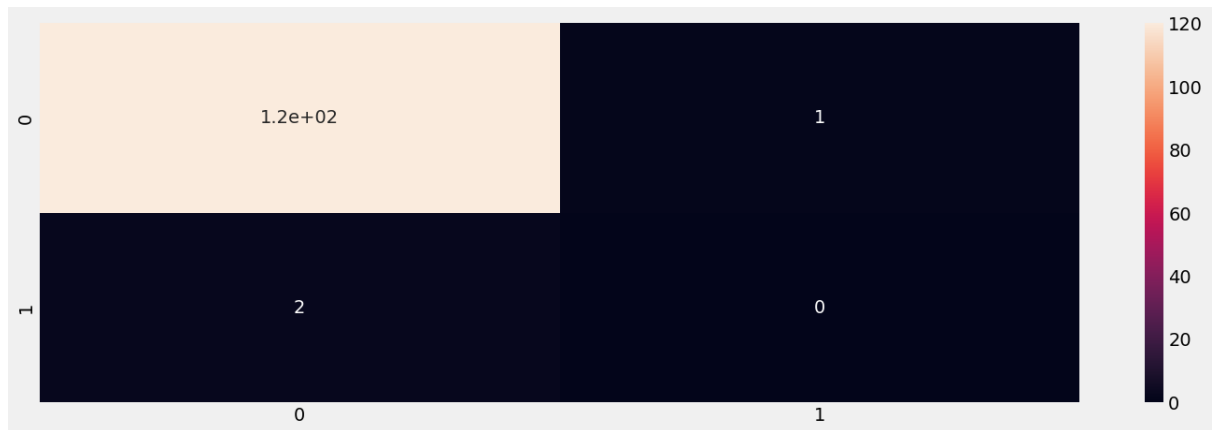


Figure 9: Confusion matrix for GB

- Support Vector Machine (SVM): Splitting ratio was 3:1. SVC from sklearn.svm.

Accuracy was 100%. Confusion matrix:

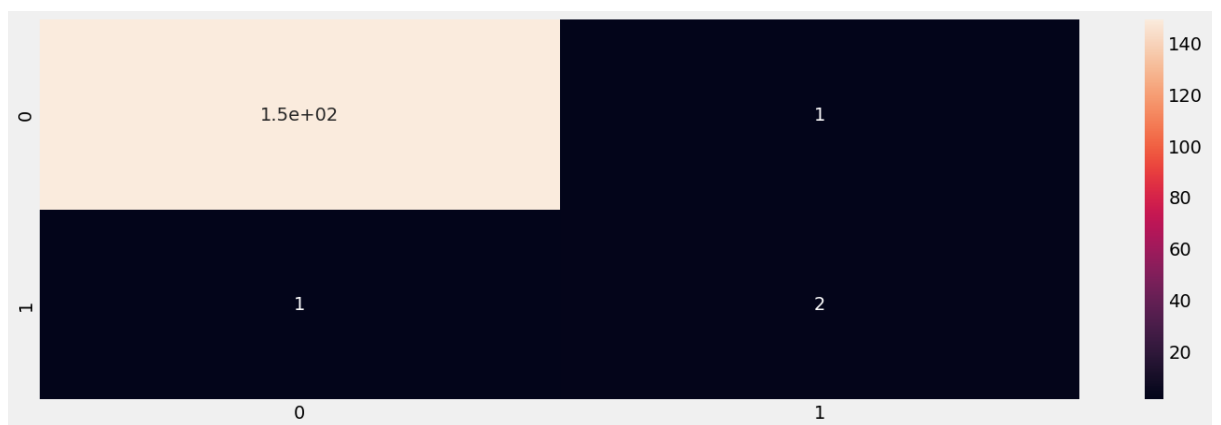


Figure 10: Confusion matrix for SVM

- KNN: Splitting ratio was 3:1. KNeighborsClassifier from sklearn.neighbors.

Accuracy was 100%. Confusion matrix:

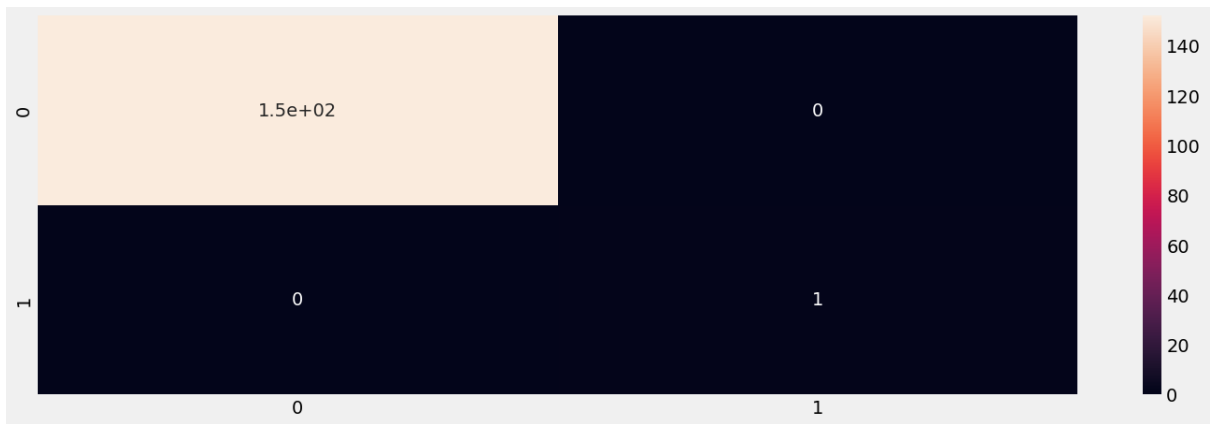


Figure 11: Confusion matrix for KNN

- Naive Bayes (NB): Splitting ratio was 3:1. MultinomialNB from sklearn.naive_bayes.

Accuracy was 98.6%. Confusion matrix:

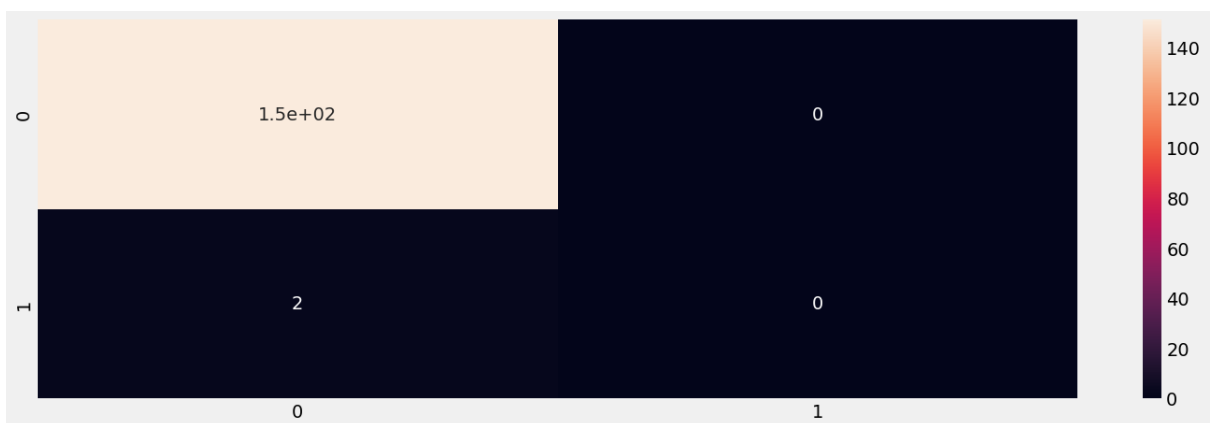


Figure 12: Confusion matrix for NM

3.5 Data Pre-processing

The implemented 8 algorithms can be compared with each other from multiple points of views. The confusion matrix provides various indications of the algorithm's performance. For e.g. the precision of the algorithm, which is the ratio TP to (TP + FP) tells us how many of the positive predictions made by the model were actually correct.

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

Recall/sensitivity of the algorithm tells us how many of the positive instances were correctly predicted.

$$Sensitivity = \frac{TP}{TP + FN}$$

F1_score which is the harmonic mean of precision and sensitivity is useful when working with an imbalanced dataset.

$$F1Score = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity}$$

Lastly, we can calculate precision, recall and F1_score for individual categories of the target class and calculate the average and weighted average of those accuracy measuring metrics.

In this section we will compare and analyze accuracy score and classification report (precision, sensitivity, f1_score etc) of the 8 implemented algorithms. We begin with comparing accuracy scores (validation accuracy scores).



Figure 13: Comparison of accuracy score of all the algorithms

The lowest accuracy was for the Decision Tree algorithm, 97.6%. Random Forest Classifier, AdaBoost and KNN, all three algorithms had the highest accuracy of 100%.

Now we compare the precision, sensitivity and f1_score in tables.

- Precision:

	LR	DT	RF	Ada	GB	SVM	KNN	NB
0	1	1	1	1	0.98	0.99	1	0.99
1	0.5	0.25	1	1	0	0.67	1	0
Avg	0.75	0.62	1	1	0.49	0.83	1	0.49
Weighted-Avg	1	0.99	1	1	0.97	0.99	1	0.97

Table 3: Precision of all 8 algorithms, splitted by the category of the target class

- Sensitivity:

	LR	DT	RF	Ada	GB	SVM	KNN	NB
0	0.99	0.98	1	1	1	0.99	1	1
1	1	1	1	1	0	0.67	1	0
Avg	1	0.99	1	1	0.5	0.83	1	0.5
Weighted-Avg	0.99	0.98	1	1	0.98	0.99	1	0.99

Table 4: Sensitivity of all 8 algorithms, split by the category of the target class

- F1_Score:

	LR	DT	RF	Ada	GB	SVM	KNN	NB
0	1	0.99	1	1	0.99	0.99	1	0.99
1	0.67	0.4	1	1	0	0.67	1	0
Avg	0.83	0.69	1	1	0.5	0.83	1	0.5
Weighted-Avg	0.99	0.98	1	1	0.98	0.99	1	0.98

Table 5: F1_Score of all 8 algorithms, split by the category of the target class

Similar to accuracy score, we also observe here that AdaBoost, Random Forest and KNN had the highest scores (100%)

Conclusion

With the use of machine learning algorithms and statistical models, this research effectively created a predictive model which can detect depression among university students in Bangladesh. The predictive model has good accuracy scores on its training dataset. The dataset was created via surveys, the survey was distributed to many university students in Bangladesh. The main goal of our research is to improve student well-fare and tackle the underlying factors of depression. The youth in our country are vulnerable to this problem. So, it is crucial to tackle depression and gear up focus on university students' mental health.

References

- [1] Hannah Roberts, Marco Helbich, Julian Hagenauer, "Relative importance of perceived physical and social neighborhood characteristics for depression: a machine learning approach," *Social Psychiatry and Psychiatric Epidemiology*, 2019.
- [2] Chong, Sunil Kumar and Ilyoung, "Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States," *International Journal of Environmental Research and Public Health*, 2018.
- [3] *A temporal analysis of depression related tweets*, 2022.
- [4] Md Shahriar Rahman Rana and Md Rayhan Kabir, "Determining Clinical Depression From The Analysis of Socio-Economic Attributes," *IEEE*, 2021.
- [5] Yingfei Zhang and Zheng Fei Ma , "Impact of the COVID-19 Pandemic on Mental Health and Quality of Life among Local Residents in Liaoning Province, China: A Cross-Sectional Study," *International Journal of Environmental Research and Public Health*, 2020.
- [6] Nehshad Hosseinifard, Mohammad Hassan Moradi and Reza Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal," *Pubmed*, 2012.
- [7] Rouzbeh Razavi, Amin Gharipour and Mojgan Gharipour, "Depression screening using mobile phone usage metadata: a machine learning approach," 2020.

- [8] Arkaprabha Sau and Ishita Bhakta, "Predicting anxiety and depression in elderly patients using machine learning technology," *Creative Commons*, 2017.
- [9] Umme Marzia Haque ,Enamul Kabir, Rasheda Khanam, "Detection of child depression using machine learning methods," *Plos One*, 2021.
- [10] László A. Jeni, Jeffrey F. Cohn, Fernando De La Torre, "Facing Imbalanced Data - Recommendations for the Use of Performance Metrics," *IEEE*, 2013.
- [11] Kerry A. Schwanz, Linda J. Palm, Samuel F. Broughton, Crystal R. Hill-Chapman, "Self-Reliance and Relations with Parents as Predictors of Anxiety and Depression in College Students," *Research in Psychology and Behavioral Sciences*, 2016.
- [12] Rumana Rois, Manik Ray, Atikur Rahman & Swapan K. Roy , "Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms," *Journal of Health Population and Nutrition*, 2021.
- [13] Ahnaf Atef Choudhury, Md. Rezwan Hassan Khan, Nabuat Zaman Nahim, Sadid Rafsun Tulon, Samiul Islam, Amitabha Chakrabarty, "Predicting Depression in Bangladeshi Undergraduates using Machine Learning," *IEEE*, 2019.
- [14] Sunil Kumar andIlyoung Chong, "Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States," *Int. J. Environ. Res. Public Health* , 2018.
- [15] Haomiao Jin, Shinyi Wu, "Developing Depression Symptoms Prediction Models to Improve Depression Care," Big Data and Health Analytics Conference, 2014.

- [16] Abdullah Al Taawab, Mahfuzzur Rahman, Zawadul Islam, Nafisa Mustari, Shaily Roy, Md. Golam Rabiul Alam, "Detecting Self-Esteem Level and Depressive Indication Due to Different Parenting Style Using Supervised Learning Techniques," *IEEE*, 2022.
- [17] Jasiya Fairiz Raisa, M. Shamim Kaiser & Mufti Mahmud, "A Machine Learning Approach for Early Detection of Postpartum Depression in Bangladesh," 2022.
- [18] V. V. Sankaranarayanan, J. Sattar and L. S. Lakshmanan, "Auto-play: A data mining approach to ODI cricket simulation and prediction.," in *SIAM International Conference on Data Mining*, 2014.
- [19] S. Brian, "The Problem of Shot Selection in Basketball," *PLoS One*, 25 January 2012.
- [20] T. Tulabandhula and C. Rudin, "Tire Changes, Fresh Air, and Yellow Flags: Challenges in Predictive Analytics for Professional Racing.," *Big data*, 2014.
- [21] R. D. Choudhury and P. Bhargava, "Use of Artificial Neural Networks for Predicting the Outcome," *International Journal of Sports Science and Engineering*, vol. 1, no. 2, pp. 87-96, 2007.
- [22] F. Duckworth and T. Lewis, *Your Comprehensive Guide to the Duckworth/Lewis Method for Resetting Targets in One-day Cricket*, University of the West of England, 1999.
- [23] I. Bhandari, E. Colet, J. Parker, Z. Pines, R. Pratap and K. Ramanujam, "Advanced Scout: Data Mining and Knowledge Discovery in NBA Data," *Data Mining and Knowledge Discovery*, pp. 121-125, March 1997.

- [24] Arkaprabha Sau and Ishita Bhakta, "Predicting anxiety and depression in elderly patients using machine learning technology," *Creative Commons*, 2017.