Identification and Comparison of Factors Behind Mental Turbulence of Public and Private University Students

by

Md. Akil Tahsin
ID- 17101510
Sumaiya Hannan
ID- 17101384
Fabiha Rumman Shifa
ID- 21141042

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
June 2021

Declaration

It is hereby declared that

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

Student's Full Name & Signature:

Md. Akil Tahsin
ID- 17101510

Sumaiya Hannan
ID- 17101384

Fabiha Rumman Shifa
ID- 21141042

Approval

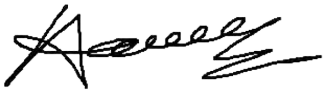
The thesis titled "Identification and Comparison of Factors Behind Mental Turbulence of Public and Private University Students" submitted by

1. Md. Akil Tahsin (ID- 17101510)
2. Sumaiya Hannan (ID- 17101384)
3. Fabiha Rumman Shifa (ID- 21141042)

Of Spring, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on June 10, 2021.

Examining Committee:

Supervisor:
(Member)

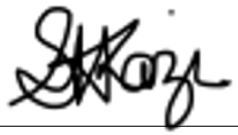


Dr. Amitabha Chakrabarty
Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

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

Head of Department:
(Chair)



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

Ethics Statement

Collecting data set related to mental health, emotions, feelings and subjective opinions are sensitive topics. Therefore, we did not collect any personal data such as name, phone number or any other personal identification criteria. All the data collected are anonymous for the security and reassurance of every participant.

Abstract

Among the young generation of Bangladesh, especially university students, mental health is a challenging but ignored topic. The primary goal of this paper is to analyze mental states found in university students all while trying to find out comparative distinction between factors prominent in public and private university environment. Professional consultation has been taken to determine probable factors related to a student's university life and mental health. The dataset is split into public and private university students to map the specific involvement of aforementioned factors. SMOTE has been used to handle data imbalance for minority classes. Machine learning algorithms such as Logistic Regression, K-Nearest Neighbour Regression and Random Forest Regression have been evaluated with MAE and R2. The most optimal regression algorithm followed by Apriori data mining algorithm have been discussed to build a hybrid model. Most influential factors are identified, comparative analysis of selected features' influence levels and associations are explored and probable causes are discussed.

Keywords: Data Mining; Machine Learning; Frequent Feature; Association Rule; Apriori; Logistic Regression; KNN; K-Nearest Neighbour; Random Forest Regression; SMOTE; Synthetic Minority Oversampling Technique; MAE; R-Squared; Optimization; Identification; Depression; Mental Health; University; Student; Education; Bangladesh; Public University; Private University; Society; Social Impact; PHQ-9

Dedication

We would like to dedicate this research to our parents who supported us all throughout our lives and guided us to become better human beings.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption. Secondly, to our thesis supervisor Dr. Amitabha Chakrabarty sir who provided us with insights, suggestions on our thesis topic and how we should approach the problem statement. Also, to our thesis co-supervisor Mr. Saiful Islam sir for his kind support and advice in our work, as well as providing necessary resources and background study materials that guided our research. He helped us whenever we needed help. Thirdly, to all the classmates, seniors, juniors who helped with their precious time to provide support and well wishes as well as their valuable input in our data collection. Fourthly, to our parents. Without their throughout mental support it may not have been possible. With their kind guidance and prayers we are now on the verge of our graduation. Finally, we would like to acknowledge our thesis group mates. Without our dedication, teamwork and supportive mentality, we would not have been able to complete this research.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$ADASYN$ Adaptive Synthetic Sampling

$COVID$ Corona Virus Disease

$DSM - IV$ Diagnostic and Statistical Manual of Mental Disorders - edition 4

etc etcetera

KDE Kernel Density Estimate

KNN K-Nearest Neighbors

MAE Mean Absolute Error

MHP Mental Health Professional

MSE Mean Squared Error

PHQ Patient Health Questionnaire

$PTSD$ Post-traumatic stress disorder

REG Regression (plot)

$SMOTE$ Synthetic Minority Over-sampling Technique

$USDI$ University Student Depression Inventory

WHO World Health Organization

Chapter 1

Introduction

1.1 Overview

Depression is a mental state. It can happen to anyone. And most of the time external image of any human being does not truly reflect the state of mind. People going through any sort of mental problem can suffer from depression, anxiety and many other kind of health diseases. In this modern era of technological and medical advancement, identifying mental issues and spreading awareness on how to deal with it has become a global cry. Now, more than ever, the global populace is gradually paying attention to the mental health and acknowledging how much effect it has on a person's regular life.

However, such is not the scenario in Bangladesh. This developing country, having a huge part of the population comprising of students, do not acknowledge mental health issues as much as it should. As the world stands right now, it is a necessity for a country to focus on every individual's mental health, in order to ensure a healthy and sound population. Nevertheless, the mental health of students in this country is, more often than none, neglected. Students from kindergarten to undergraduate life and sometimes even post-graduate students in this country face different type of social stigma that disturbs the state of their mind and disrupts the mental growth. Especially the undergraduate students are more likely to undergo mental health issues due to the nature of this country's social structure.

University, in the most basic definition, stands for, "A place of higher education usually for people who have finished twelve years of schooling and where they can obtain more knowledge and skills, and get a degree to recognize this." [23]. As such, the universities of this country are divided into two major sanctions: public and private. Public universities are funded by the Government while private universities rely on tuition and funding from private individuals or organizations [31]. In truth, there also exists, albeit very few in numbers, some universities that are maintained through international organizations and thus have the stature of being an "international" university. Regardless, both international and private universities are referred as the same (private university) due to their nature of fee structure and financial benefits.

Whether a student belongs to a public university or a private institution makes a big difference in their social lifestyle as well as how society perceives them. The state of one's university directly or indirectly plays a roll on how they are treated, what expectations they have to fulfill and what type of responsibilities they are bestowed

upon. Due to all these, a student, be that from a public university background or a private university affiliation, can and does indeed face different sorts of mental instability and issues. A student can fall into depression just by the different factors that are present in their own university as well as factors that are absent therein.

1.2 Importance of the Research

Depression has caused an alarming situation as the rate of suicide due to depression is increasing. People who have had a traumatic experience in their lives are more likely to develop depression. Depression can do severe damage to one's health, especially when it lasts for a long time and has a moderate or severe intensity. In recent days, during the pandemic period, some suicide cases of young people have been witnessed which happened due to depression. In Bangladesh, almost 7 million people suffer from depression [22]. However, the causes of depression are not very clear yet.

One of the major challenges in depression in Bangladesh is that, there are lack of studies. If we do not understand what is influencing the young minds towards depression, we cannot find a proper way to treat them. Unfortunately, no proper research was found on after COVID-19 related mental issues in Bangladesh yet [18].

In Bangladesh, university students fall in depression for various reasons. Identifying them has become an immediate need as early identification might help one to come out of depression with proper help. However, to identify it easily, it is needed to categorize the reasons based on different situations. In Bangladesh,the distinction between public and private universities has been an issue for the undergraduate students for a long time. Social pressure, facility issues, graduation time etc. factors work here. If we can understand how much these factors are affecting individuals, we might be able to recognize their depression level as well. Living with depression is difficult, therefore identifying and taking steps in the early stage is the most important thing.

To conclude, the importance of this research is immense as we find out the difference of depression level and reasons of the young people, who are currently studying in universities; differentiating them based on whether they are studying in public university of private university.

1.3 Current Scenario and Motivation

According to WHO, Depression is a global mental disorder affecting more than 264 million people in the planet [25]. Depression can be dangerous to one's health, especially if it lasts for a long time. Despite the fact that there are effective therapies for mental diseases, between 76 to 85 percent of people in low and middle-income regions do not get proper treatment. Lack of funding for the research of mental disorders, a shortage of skilled health-care practitioners who can help people when they are in the first stage of their mental turbulence and the societal stigma associated with mental illnesses are the primary obstacles to effective treatment.

If we look into the current scenario, we can see that the rate of depression in young people of Bangladesh is increasing rapidly. In a recent study, 33.3% depression was identified from a nationwide study in Bangladesh [18]. In that study, it was seen that most of the depressed people were of young age. Depression can result in stress and dysfunction, deteriorating the affected person's circumstance as well as the depression itself. Therefore, it is very important right now to do proper research on this. This motivated us to work on this research more so that we can contribute something to our society. We believe, if proper study and research is done in this area, it will be easier to identify depression in the early stage for the person and his/her surrounding people.

1.4 Research Objectives

In the context of Bangladesh, undergraduate time is very significant in a student's life as they train themselves for a desired career during these four or five years. Most students start their undergraduate life thinking it as a bed of roses. With time, they understand the real world and start feeling the pressure from different sources. Such as, family pressure to succeed, faculty and peer pressure to do good, sometimes pressure from their partner to get a good job soon. All these pressures could lead one to different types of mental illnesses; for example, anxiety issue, depression. Prior researches have shown that anxiety, depression etc. impact one's entire lifestyle badly which puts them under more pressure. Unfortunately, in Bangladesh the mental health has been neglected for long. [20] In total, there are 53 public universities and 80 private universities in Bangladesh. Nevertheless, there is unfortunately no model to track their degree of mental instability.

Depression is viewed as a genuine ailment that can deteriorate without appropriate treatment. The individuals who look for treatment regularly observe upgrades in manifestations in only half a month [28]. It is a serious mental illness that adversely influences how one feels, acts and thinks [33]. Depression generally starts in the teenagers, 20s or 30s, however it can occur at whatever stage in life [26]. Our research objectives are -

1. To measure the mental turbulence among undergraduate students in both public and private universities by using PHQ-9.

2. Secondly, to record the data factors surrounding students' lives that may have an effect on their mental health.

3. And lastly using machine learning and data mining algorithm to identify the importance and frequency of those data set factors, in order to try and find correlation between the factors, the university background and students depression states.

With the results in hand, we can suggest useful measurements to improve their mental condition. Which will support the students, first and foremost, and the nation overall. Our undergraduate students are the nearest to leaving a mark in the country's future. Hence, it is very necessary to maintain their mental health.

1.5 Challenges Faced

First of all, we discussed and decided upon our research topic after much careful consideration regarding the mental health issues in current students. After finalising the research topic, our first and foremost challenge was to collect data. Research on students' mental health with respect to their university background have not been conducted in this country before. So, it was a big challenge to collect dataset with large enough data that could provide us an accurate result.
Then, we had to take appointment of professional mental health experts and psychiatrists. With extensive discussion and follow-up visits to their workplaces, we were finally able to make a valid questionnaire to conduct survey for our dataset.
Conducting the research during COVID-19 pandemic was the biggest hill that we had to climb. Not to mention, our research requires data from students of private and public universities to get information about their mental condition, personal life and university aspects. To do so, our initial plan was to visit a number of universities all around the country and physically take surveys for data collection. But before implementing any of the ideas, all the universities closed off due to COVID-19. So naturally, we were short in information. We had to use online platform as our only source of data collection. We had to personally communicate with a lot of friends, classmates and acquaintances studying in different universities and pass our survey through them.
Even after collecting enough data samples for primary research, we found imbalance in our dataset. Therefore, we needed to perform appropriate data balancing techniques to introduce balance in the dataset. Finally we managed to prepare a potential dataset to fit in the model.

1.6 Thesis Outline

The primary goal of this research was to develop a model which can precisely identify the prominent factors from a university student's surroundings that may have influence over his/her mental turbulence. Our intention was to analyze a large data of both public and private university students and find out the strong correlation of these factors with their mental health. We also wished to have a comparative analysis between factors that are prominent in public universities and factors that are more frequent in private universities. Our aim was to build a model with high accuracy and estimation capabilities. Furthermore, we have tried to visualize as much data we could.

First of all, in (Chapter 1) a detailed overview of mental health in Bangladesh was provided. Current scenarios in this country regarding mental health awareness were discussed. A three-fold plan of data collection, factor identification and correlation evaluation as our research objective was described.

Secondly, in (Chapter 2) theoretical knowledge and global research on mental health were explored. Secondary research was conducted to make hypothesis on probable depression causes.

After that, in (Chapter 3) Background theoretical knowledge on intended algorithms

were visually presented and described. The chapter consisted of concepts of models that will be implemented later.

Then, in (Chapter 4) methodology of this research and the workflow was proposed. Discussions were made regarding our data and processing techniques to convert the data viable for algorithm implementation. Proper data visualizations took place. Necessary arrangements were made to conduct experimental analysis.

After that, in (Chapter 5), machine learning algorithms were implemented and evaluated. Selected algorithms were used to generate frequently occurring important factors.

Next, in (Chapter 6) Results were analyzed. An in-depth discussion was made regarding probable causes and effects of prominent features. Comparative analysis between public and private university factors was presented.

Finally, in (Chapter 7), conclusion was drawn and further work plans were elaborated.

Chapter 2

Literature Review

Even though mental health is a prevalent issue in the current generation, there has not been much research explicitly on the mental health issues in Bangladeshi undergraduate university students. There are not many data sets or relevant work found on the factors of Government funded and privately owned universities and their correlation to the mental issues of these universities' students. Therefore, much work of this research relied on primary data collection and field research and queries with specialists in the country.

2.1 Depression

Depression is a mood disturbance that induces a persistent sense of distress and lack of interest. As per [26] it affects a person's behavior, thought and action and can lead to a number of mental and physical difficulties. It is often called major depressive disorder or clinical depression. It can cause one to face problems doing regular day-to-day things, and it may make them question the purpose of living sometimes. Depression is not a weakness. It is difficult to snap out of it. One may need long term therapy to cure this mental condition. It might sound hectic or scary. But it is very much possible to recover from depression with medicine, psychotherapy or both for most individuals.
Moreover, [16] tells that depression is a growing psychological disorder many physicians undergo. This impairs not only the standard of living greatly, it also sometimes complicates underlying medical problems like diabetes. It is vital for all patients with chronic medical conditions to consider the symptoms of depression. Once diagnosed, the psychiatrist has a major part to perform in communicating to the patient, the existence and diagnosis of depression. Antidepressant medications and/or psychiatric therapy are essential. It is also necessary to evaluate suicidal risk when assessing distressed patients. The need for referral to a doctor suggests complicated conditions, inability to respond to the initial care or worries about the possibility of suicide.

2.2 How depression is misunderstood

As mental illness has come to light since recent few years, misconception about depression has become very common within people. Some misunderstanding re-

garding depression can be learnt from [11]. Depressed people feel that nothing can make them happy anymore. It is not just a sad feeling rather it is much more than that. Most people assume that they are depressed when they are feeling down. Secondly, people who think negatively are often misunderstood as depressed. It is true that depression forces people to feel negative emotions. Because Depression is a hormonal imbalance and it causes less production of dopamine, serotonin, etc.(happy hormones). Depressed people can not simply get over negative feelings. But that does not imply that all people who think negatively are suffering from depression. Another misconception is that people do not feel depressed about particular things. A person can live an amazing life and still be depressed. Next, depression is not always triggered by a specific event. People tend to assume that their mental illness is probably caused by the most traumatic event in their life. But that is not the case all the time. In most of the cases people suffer from depression because it has been running in their family. Lastly, depression is not something people can easily get over with. Depression can be extremely difficult to cope with, between the difficult feelings and the uncertain nature of the condition. Depressed people tend to be very hard on themselves. Depressed people put themselves down to the level of feeling completely insufficient, which just causes them to spiral into depression more and more.

2.3 Related works

The common causes of depression among the students is discussed by the author of [7]. Some of our findings from this article are (i) Socio-economic status is the first significant trigger of depression in students. A widespread international study (from 23 countries) on the link among depressive symptoms and socioeconomic history of university students (Steptoe, Tsuda, Tanaka and Wardle, 2007) found that student depression was due to family and personal incomes, parental employment and family property. Similarly, a review of Egyptian research results indicates that socio-economic conditions are adversely related to the effects of depression in adolescents. This research found that students from low-income households and parents appear to become distressed.(Ibrahim et al., 2013). (ii) Students from the rural areas have higher distress rates than students residing in metropolitan regions (Bayram and Bilgel, 2008). Shamsuddin and his colleagues (2013) said it could be attributed to an economic condition in which families in rural areas generally have a lower economic standing. Furthermore, some students from rural areas may migrate away from the university and neglect their families which may create some problems for some students. (iii) Studies often indicate that students from higher socio-economic groups are more likely to feel in charge. This feeling of responsibility will help shield students from mental health issues such as depression in the university community (Lanchman & Weaver 1998). (vi) In addition, students' education level may be a reason behind their mental turbulence. (v) Another important element is living away from home or moving to a new setting like a workplace. Some students can experience psychological distress, particularly depression, if separated from home or family. This is attributed to challenges and difficulties linked to life in a diverse and unfamiliar world without social help at a university or college (Asyan, Thompson & Hamarat, 2001). Increased pressure and tension regarding university have also been shown to have less effect on students who stay home with their family (Christie,

Munro & Retting, 2002). (vi) Other than that, another issue with housing is that students who live alone, particularly foreign students, may face problems without social help. A research by Haldorsen et al.(2014) showed that the signs of depression are greater than that of the students who work for their families in the case of social difficulties. The following research indicates that students from university and college have a better likelihood of surviving without depression while students stay with their families and spouse and provide social help for resolving their problems. (vii) In a recent Haldorsen et al. (2014) report, an significant observation was found that students' stress conditions are significantly correlated with depressive symptoms. Haldorsen et al (2014) hypothesized stated in the previous research that higher tension in students contributed to higher depression rates. Likewise, the association between depression, tension and anxiety was illustrated by Bayram and Bilgel (2008). (viii) Students' satisfaction is the third main parameter of depression. This suggests students who don't have a higher risk of dissatisfaction than others or who are content with their studies (Bayram & Bilgel, 2008). The student's loss of confidence and enthusiasm may be a potential explanation of this observation, as parents of the student often chose a topic for research (Chen et al., 2013).

The study [3] involved 820 students (54.3% male and 45.7% female), and the mean age was 22.3 years. The findings indicate a frequency of extreme depression of 7.0 percent and mild to medium depression of 25.2 percent. The loss of social interaction was correlated with multivariate logistic regression, with successful PTSD evaluation and with a mild to severe sleep crisis. High depression levels have been reported. Several risk factors were reported including co-morbidity (PTSD and difficulty sleeping) and absence of social support.

The incidence of depression is slightly higher in non-athlete male undergraduates than in male athletes [5]. The findings also indicate that female students had a significantly higher degree of depression than males. This may be induced by the tiredness and energy deficiencies that are more prevalent among women relative to males. The degree of depression among male and female graduates was unfavorable for physical activity. There is, however, no clear correlation between physical exercise and depression level based upon the age of male and female athletes.

Self-esteem and superiority are the two most significant mediators of the family adversity connection to depression. Yet self-esteem and parental encouragement are the other two mediators of the relation of depression to personal adversity. As described in [8]. Therefore, the cumulative mediating impact of wealth for family hardship is stronger than for personal hardship. These findings suggest that the pathways involved in the transmission of stress to distress differ considerably between specifically perceived hardship and indirect family hardship.

The article [13] says community help deficiency is a recognized determinant to mental well being disorders. Parents and friends' social interaction has become a significant indicator of depressive symptoms. The social reinforcement of family and acquaintances has been an important predictor of the quality of life (psychological). Financial encouragement from significant partners and associates also estimated the standard of life (private relationships). Public assistance networks are a powerful

outlet for the emotional well being of students in universities.

A research [4] revealed that (i) single students in comparison to married students were vulnerable to depression. This may be attributed to a difficult condition with adult students (e.g. jobs, culture, graduation and marriage). In comparison to this research, there were reports that indicated higher rates of depression in married students (ii) Throughout our research, no gender disparities throughout depression. According to our results, prior research has found that there were no variations in depression amongst students. It may result from the reality that the same stresses extend to Iranian female university students. However, several findings are contradictory to our results and indicate that female students have higher rates of depression. (iii) With unique intellectual, financial and interpersonal stresses, the University is an essential stage of transient existence. Due to these changes, the likelihood of depression can be increased. Nevertheless, in the present research, the incidence of anxious symptoms is more pronounced than the norm.

We learnt the symptoms of depression from [27]. As everybody, it states that young people will often feel irritable, often feel irritable and are especially vulnerable to rejection and criticism. Yet the young adult might have depression if such moods continued for two weeks or more. Symptoms that may suggest depressants include: becoming irritable or grumpy with exhausted feelings of worthlessness or remorse much of the time while thinking about death or suicide sleeping – either falling asleep or sleeping without reasons, and realizing it is hard to lose motivation or lose weight or get weight from tobacco, alcoholic drinks or food. There are often no obvious symptoms of depression, however in young adults, parents can notice behavioural improvement that indicates stress and can not be overlooked. These include: lower scores for the emotional isolation through the usage of alcohol and narcotics through classrooms.

Chapter 3

Background Analysis

3.1 Patient Health Questionnaire (PHQ-9)

PHQ-9 is a patient questionnaire that can be used very easily [30]. This is a self-administered edition of the popular mental illness diagnosis tool PRIME-MD. This model determines depression level based on the scores of nine DSM-IV criteria as '0' (not at all) to '3' (nearly every day). PHQ-9 is a verified model to use in primary care, as referred by Dr. Yasir Arafat [35]. PHQ-9 is not a diagnostic instrument, but is used to assess the severity of depression and treatment approaches. However, in at-risk patients - eg, those with coronary heart disease or after stroke - it may be used to make a preliminary diagnosis of depression.

[1] There was a significant decrease in functional status on all 6 SF-20 sub scales as PHQ-9 depression intensity increased. Symptom-related challenges, sick days, and the use of health care have also increased. A PHQ-9 score of more than or equal 10 had a sensitivity of 88% and a precision of 88% for major depression using the MHP re-interview as the criterion model. The 5, 10, 15, and 20 of PHQ-9 scores reflected mild, moderate, moderately severe, and severe depression, respectively. In the primary care and obstetrics-gynecology studies, the findings were close.

In figure 3.1 we observe the PHQ-9 questionnaire and each question's assigned values in a range of 0 to 3.

Over the last *2 weeks*, how often have you been bothered by any of the following problems?
(use "✓" to indicate your answer)

	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 figety 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	0	1	2	3

Figure 3.1: PHQ-9 Questionnaire

3.2 Decision Tree

While applied in a model, Decision Tree intakes some features of the data and predicts the target based on that. Decision Tree speculates the data and distributes them into groups or classes using data processing. Classification problems can be easily shown with this algorithm. Let's say for instance, all the features are labeled and they have a finite amount of values. One of the labels is the target that the algorithm has to predict. Decision Tree distributes all the labels into nodes. That is why this algorithm is also known as Classification Tree. The arcs that originate from a node labeled with an input feature are either labeled with each of the target feature's potential values, or the arc goes to a subordinate decision node on a different input feature. Every leaf node is a class. A class in a decision tree is generated from each of the entries in the data.
In a Decision Tree, the leaf nods are the children of the root node. This structure is achieved by splitting the entire data. In every subset, the node is classified according to its features in a recursive manner. This process stops when all the leaf nodes reach the target that was supposed to be predicted[24].
Data are recorded in the following form:

$$(x, Y) = (x, x, x, x, \ldots, x, Y) \tag{3.1}$$

The target that has to be predicted is denoted by Y. While x is a vector incorporating all the features or labels. Figure 3.3 shows a visual representation of a Decision Tree.

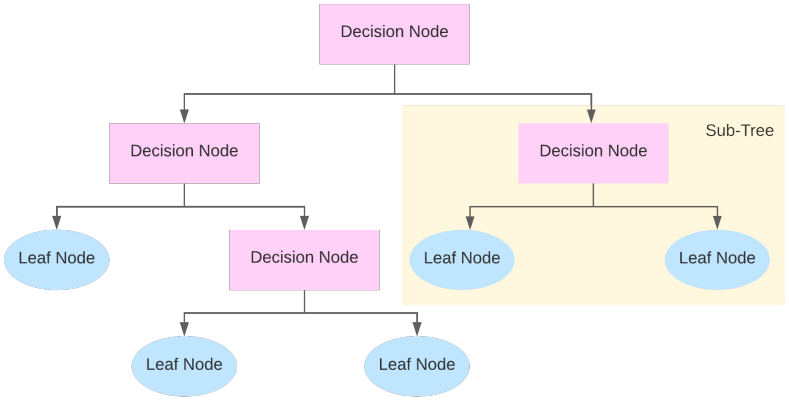


Figure 3.2: Decision Tree Infograph

3.3 Random Forest

For our supervised learning based research purpose, Random Forest Regressor has been used. But this algorithm is useful to solve classification problems as well.
Random Forest is a widely used machine learning algorithm. It is a supervised problem because it deals with labels in a dataset. It predicts the target based on the input features. After fitting the dependent variable Y (target) and the independent variables x (labels), Random Forest generates the desired amount (n) of decision trees in a random way. Every decision tree predicts an outcome. Then Random

Forest generates a decision based on the majority outputs from the n number of the trees. The number of trees and random states value can be tuned to reach the desired goal [21].

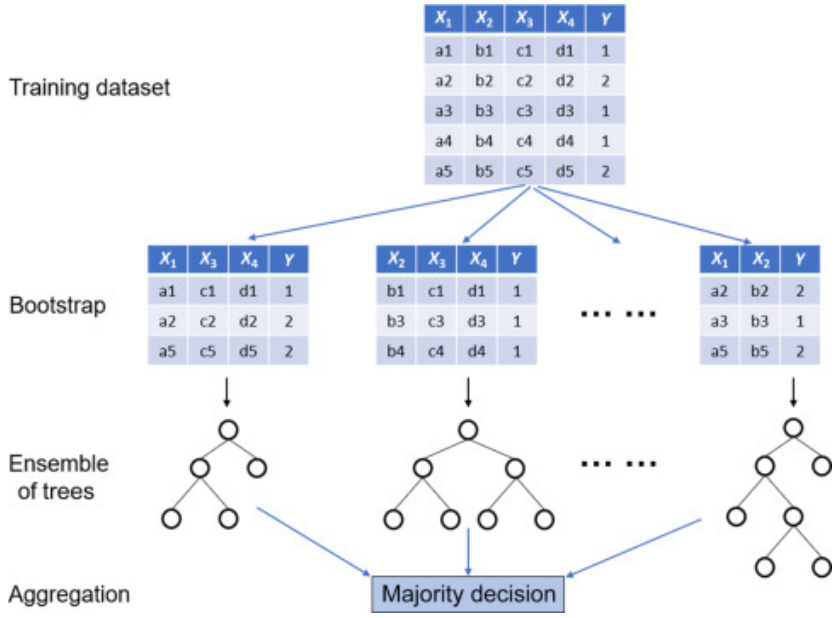


Figure 3.3: Random Forest Demonstration

The technique for splitting a node only considers a random subset of the features. Rather than searching for the best possible thresholds, it is feasible to make trees more random by employing random thresholds for each feature (like a normal decision tree does) [32].

Another nice usefulness of this ensemble algorithm is that determining the relative value of each feature on the forecast is quite simple. Sklearn has an excellent tool for measuring the relevance of a feature by looking at how much the tree nodes that use it reduce impurity over the entire forest. After training, it calculates this score for each characteristic and weighs the findings so that the total importance equals one. Random Forest can play a vital role while selecting features for a model. By identifying the less important factors, it can ease down the data processing and enhance the overall speed for computation.
This algorithm determines which characteristics to exclude based on feature importance because they don't contribute enough (or occasionally none at all) to the prediction process, as shown in figure 3.4. This is significant because, in machine learning, the more features supplied, the more likely the model is to suffer from over-fitting, and vice versa.

feature	importance
Title	0.213
Sex	0.173
Age_Class	0.086
Deck	0.085
Pclass	0.071
relatives	0.070
Fare	0.069
Age	0.065
Embarkerd	0.051
Frae_Per_Person	0.045
SibSp	0.037
Parch	0.025
alone	0.011

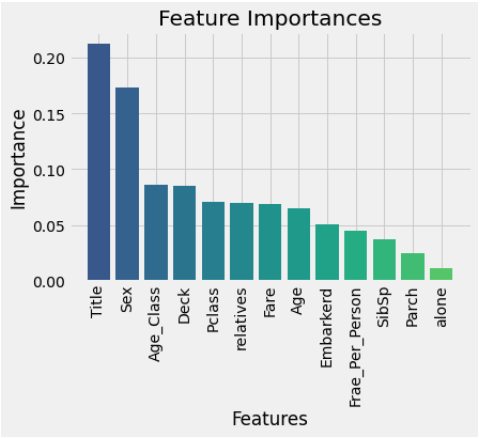


Figure 3.4: Important Feature Selection sample table and graph

3.4 Logistic Regression

Logistic Regression, as the name suggests, predicts the target using regression. Even though it is a regression model, it is often used as a classification algorithm as well. This algorithm uses a function called logistic to determine the binary or Boolean output. Which means logistic regression is useful when there is only one decision to make or the other. It's a variant of linear regression in which the target variable is categorical. The dependent variable is the log of odds. Using a logit function, logistic regression predicts the likelihood of a binary event occurring [14].

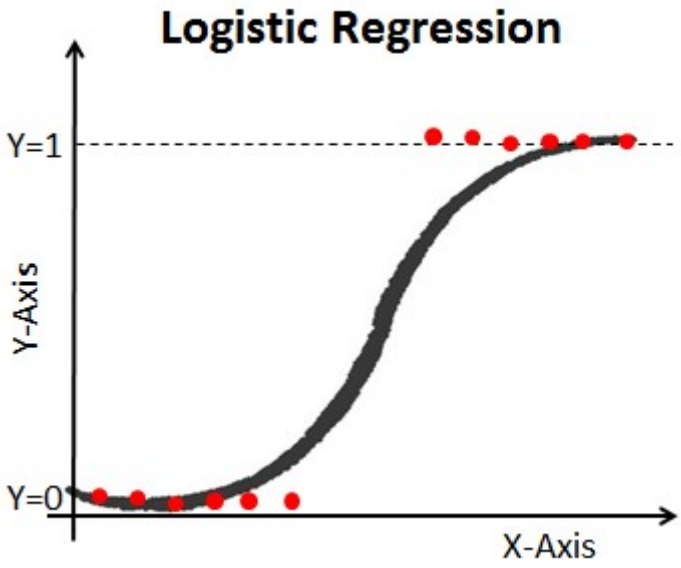


Figure 3.5: Logistic Regression

Equation of linear regression:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn$$

The dependent variable is y, and the explanatory variables are x1, x2... and Xn.
Sigmoid Function:

$$p = 1/1 + e^{-y}$$

Sigmoid function applied on linear regression:

$$p = 1/1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn)}$$

Because of its simple and efficient nature, it does not require a lot of computing power, is simple to implement, and analyze, and is extensively utilized by data analysts and scientists. It also does not necessitate feature scaling. For each observation, logistic regression generates a probability score. On the other hand, a huge number of categorical features/variables is beyond the scope of logistic regression. Overfitting can be a problem. Furthermore, logistic regression cannot tackle non-linear problems, which is why non-linear features must be transformed.

3.5 K-Nearest Neighbors

This algorithm is similar to a clustering algorithm. It identifies the labels that are closely related and creates group. Each group of items are represented with different color [10].

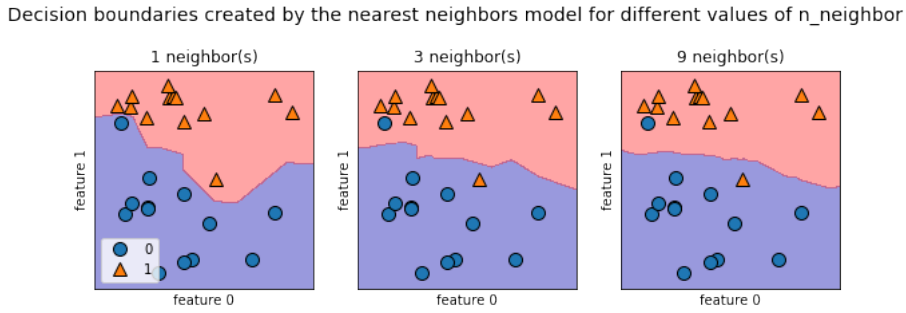


Figure 3.6: How the data points stay close to each other in KNN

KNN groups the items using a basic concept. It calculates the gap between the values and based on that it picks items that can be grouped together or that have a less gap between them. KNN is a very adaptable algorithm. It has classification, regression, and search capabilities (as we will see in the next section). It's not necessary to create a model, tweak a few parameters, or make any more assumptions. However, when samples that predict variables add up in number, the process becomes much slower [10].

3.6 Apriori Algorithm

Because it's frequently used to uncover or mine for interesting patterns and associations, Apriori is generally regarded as an unsupervised learning strategy. Apriori can also be tweaked to classify data that has been labeled.
Apriori algorithm is the most basic algorithm for locating frequently occurring objects in an item set [17]. If a set of elements meets a minimum criterion for support and confidence, it is referred to as frequent. For example, transactions involving many products purchased in a single transaction are displayed in support. Confidence depicts transactions in which products are purchased sequentially. Only transactions that match the minimal threshold support and confidence requirements are taken into account when using the frequent itemset mining method. These mining algorithms' insights provide a slew of advantages, including cost savings and a boost in competitiveness [6]. Apriori Algorithm uses association rules to create relation between the items. In a database it creates chain relations among different items, given the threshold values for support and confidence. This is known as association mining.

A possible illustration of support and confidence can be this :

$$Mobilephone => Earpod[support = 3\%, confidence - 70\%]$$

This means that 3% of customers bought mobile phones and earpods together, while 70% of customers bought the items separately. These formulas show support and confidence for Item-sets X and Y:

$$supp(X) = Number\_of\_transaction\_in\_which\_X\_appears/Total\_number\_of\_transactions$$

$$conf(X \rightarrow Y) = supp(X \cup Y)/supp(X)$$

The Apriori algorithm is a set of stages that must be followed in order to determine the most frequent itemset in a database. This data mining methodology repeats the join and prune procedures until the most frequently occurring itemset is found. The problem specifies a minimum support threshold, or the user assumes it.

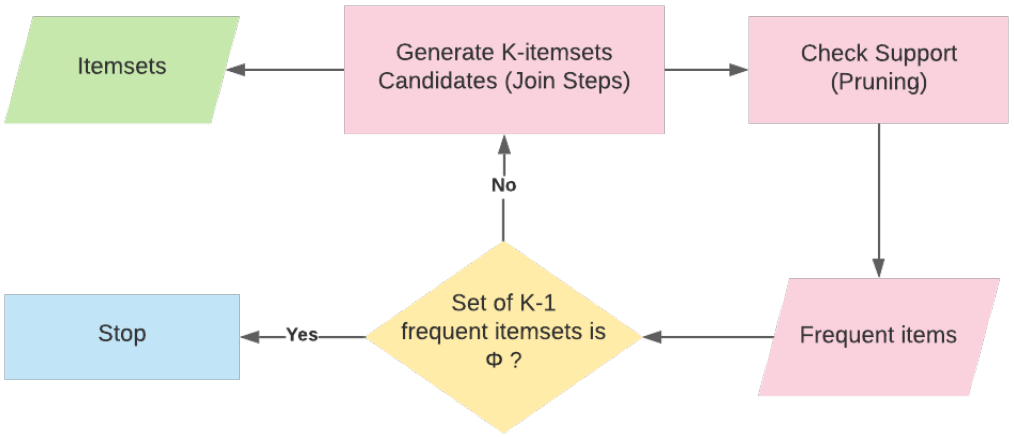


Figure 3.7: Apriori steps

Each item becomes 1-itemset of Apriori association mining's first iteration. Each itemset's occurance and support is calculated and only itemsets above the minimum threshold support is used in following iteration, other candidates are pruned. Similarly, in the next iterations, itemsets are joined within themselves and new support are calculated. In each iteration, itemsets are joined and pruned as per minimum support threshold. Super-sets are counted as 'frequent' only if all 2-itemset subsets are frequent. Such iterations will continue until most frequent itemset is reached, and then algorithm stops, as shown in figure 3.7

Chapter 4

Research Methodology

4.1 Dataset

Primary data of this research is collected via conducting surveys on students affiliated with and studying in different universities. It was planned to take physical surveys to ensure more diversified data collection and collect as much variation of both public and private university students. However, due to unfortunate circumstances and the on-going global pandemic, physical surveying measures were restricted. Hence, we collected data via online surveys through social media connectivity and tried to reach as much variety of student body as possible.

4.2 Project Workflow

As the fruit of this research we will be able to do a comparative analysis between the mental condition of public university and private university students and find out the features associated with the scenarios. The simple idea of the workflow is to build a model that will separately train on two datasets which are split based on the university type (Public or Private) from the main data set. Each data set has to be distributed among target (y) and label (X) before fitting into any model. Work strategy for this research fundamentally consists of three stages:

- Determining depression level using PHQ-9
- Using a Supervised Regression machine learning problem to filter out the important features that have influence on the PHQ-9 result
- Finding out frequently occurring features and their association with different depression levels using a Mining Algorithm

The specific steps and details of the workflow have been described in 'Work Flow Overview' in Chapter 5.

4.3 Data Pre-Processing

Initially, data was collected in categorical format. All the data inputs were in String values, and as machine learning algorithms take numerical data, it was necessary to perform data pre-processing on the dataset.

4.4 Data Encoding

Machine Learning Algorithms have varied parameters as input dataset. Some algorithms can work on categorical data where no data transformation is required. However, many machine learning algorithms cannot operate on textual or categorical data directly. They require data to be numeric. Also, there are occurrences where the datasets have numerical inputs with no specific meaning or indications. These datasets are required to be converted to generic numerical data entries. [9] These are mostly constraints that draw efficient implementation of the model.
Researches that are based on machine learning models which require numerical data inputs use several encoding methods to convert data to properly fit into the algorithm models. There are two such approaches for numerical data conversion.

1. Integer Encoding
2. 1-Hot Encoding

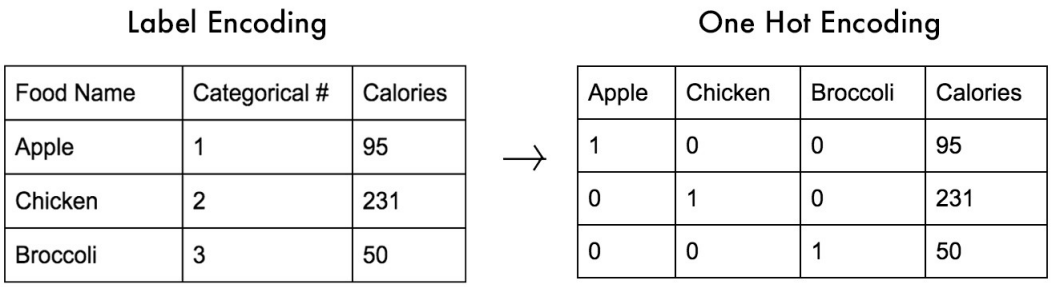


Figure 4.1: Encoding Example

Figure 4.1 shows an example of encoding. For label encoding it generates numerical values for each of the unique "Food Name", and then for 1-hot encoding, it separates each unique 'Food Name' value into a new column and assigns all the corresponding rows as '1'.

4.4.1 Integer Encoding

Also known as label encoding, this approach converts categorical data into label data where it assigns integer values to different labels in the dataset. In this dataset, we primarily performed label encoding to convert the categorical data into numerical data.
For instance, in the dataset, the column titled "Feeling tired or having little energy?" had four unique values. We renamed the column and encoded its input values as follows:

$$Not\_at\_all => 1$$

$$Several\_days => 2$$

$$More\_than\_half\_the\_days => 3$$

$$Nearly\_every\_day => 4$$

It is to be stated that, the first nine columns consist of PHQ-9 specific questions and the numerical equivalent data for each option was pre-determined [1].
In the figures 4.2 and 4.3, we can observe the before/after demonstration of a portion of our primary dataset.

Little interest or plea	Feeling down, depre	Trouble falling or sta	Feeling tired or havi	Poor appetite or ove	Feeling bad about y	Trouble concentratir
Nearly every day	Nearly every day	Nearly every day	Nearly every day	More than half the d	Nearly every day	Nearly every day
Not at all	Not at all	Not at all	Not at all	Several days	Not at all	Not at all
Nearly every day	Several days	Several days	Several days	Several days	Several days	More than half the d
Several days	Several days	Several days	Several days	Not at all	Several days	Several days
Several days	Several days	Several days	Not at all	Several days	Not at all	Several days
Not at all	Not at all	Not at all	Not at all	Not at all	Not at all	Not at all
Several days	Several days	Several days	Several days	Several days	Several days	Several days
Several days	Several days	Nearly every day	Several days	Not at all	Several days	Several days
Nearly every day	More than half the d	Nearly every day	Several days	Several days	More than half the d	Several days
Several days	Several days	Several days	Nearly every day	Several days	More than half the d	Not at all
More than half the d	Nearly every day	Nearly every day	Nearly every day	Several days	Nearly every day	Nearly every day
Nearly every day	More than half the d	Nearly every day	Nearly every day	Several days	Several days	Not at all
Nearly every day	Not at all	Several days	Several days	Not at all	Several days	Not at all
Several days	Nearly every day	Several days	Not at all	Several days	Nearly every day	Several days
Not at all	Nearly every day	Nearly every day	Nearly every day	Nearly every day	Nearly every day	More than half the d
Nearly every day	Nearly every day	Nearly every day	Nearly every day	More than half the d	Nearly every day	Nearly every day
Nearly every day	Nearly every day	More than half the d	More than half the d	Nearly every day	Nearly every day	More than half the d
Several days	Nearly every day	Several days	Several days	Not at all	Nearly every day	Several days
Several days	Several days	Nearly every day	More than half the d	Several days	More than half the d	More than half the d
Several days	More than half the d	Not at all	Not at all	Not at all	More than half the d	Not at all
Nearly every day	Nearly every day	Nearly every day	More than half the d	Several days	Nearly every day	Nearly every day
Several days	Several days	Several days	Several days	Not at all	Several days	Not at all
Several days	Nearly every day	Nearly every day	Nearly every day	Several days	More than half the d	More than half the d

Figure 4.2: Raw data

PHQ_1	PHQ_2	PHQ_3	PHQ_4	PHQ_5	PHQ_6	PHQ_7	PHQ_8	PHQ_9	Age
3	3	3	3	2	3	3	0	3	2
0	0	0	0	1	0	0	0	0	2
3	1	1	1	1	1	2	1	0	2
1	1	1	1	0	1	1	0	0	2
1	1	1	0	1	0	1	0	0	1
0	0	0	0	0	0	0	0	0	1
1	1	1	1	1	1	1	1	0	2
1	1	3	1	0	1	1	0	1	2
3	2	3	1	1	2	1	1	2	1
1	1	1	3	1	2	0	0	0	1

Figure 4.3: Processed data

4.4.2 1-Hot Encoding

For a number of machine learning algorithms, simply converted numerical values are enough. They fulfill the requirements and specific parameters of such algorithms and generate data effectively. However, other machine learning algorithms cannot work properly with any integer values. They require binary values of "0"s and "1"s to generate result and predictions.
1-Hot Encoding method takes one column and divides it into a number of columns similar to the unique values of that initial column input. Our Apriori algorithm required dataframe parameters to be 1-Hot Encoded data. As discussed in PHQ-9 section, the questionnaire generates a depression level between 1-5, which correlates to the five levels of depression; "Minimal", "Mild", "Moderate", "Moderately-Severe" "Severe"; as defined in [1].

4.5 Data Augmentation

One of the most crucial parts of a working machine-learning algorithm model is to have a considerable size of dataset to work on. However, a big obstacle for such algorithms is to have an imbalanced dataset, which refers to a scenario where there is an unequal distribution of class in a dataset. One or more particular class may have imbalanced data and the difference may lead to a decrease in the result accuracy.
There are ways to handle these sorts of imbalanced datasets, such as; undersampling and oversampling. These are some common data analysis techniques that can produce a balanced data and reduce the ratio of difference in dataset classes. [34] Oversampling is used more often than undersampling mostly due to the fact that the training data may have scarcity to begin with. Oversampling is done by transforming the minority class entries in datasets to match the number of majority class instances.

4.5.1 Synthetic Minority Over-sampling Technique (SMOTE)

In order to generate minority class instances to balance the dataset, there are several oversampling (or upsampling) techniques. Some of the most common oversampling techniques involve RandomUnderSample, ClusterCentroids, SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling). SMOTE is more useful because it generates synthetic samples for the minority class, which eliminates the problem of overfitting caused by random oversampling.
In short words, SMOTE works on the dataset by taking each minority class samples, introducing synthetic examples along the line segments created by joining the k minority class nearest neighbours. By selecting examples of minority class instances close to the feature space, drawing a line between the examples in feature space and generating new sample data points along that line, as shown in figure 4.4. SMOTE generates oversampling data to achieve desired ratio of majority class instances and minority class instances. For our research context, the desired data ratio between two classes are 1:1 [15].
In this dataset, SMOTE was performed to generate oversampling data that will augment the dataset and provide a better accuracy score for machine learning purposes. As this dataset contained imbalanced data, SMOTE was performed to balance it.

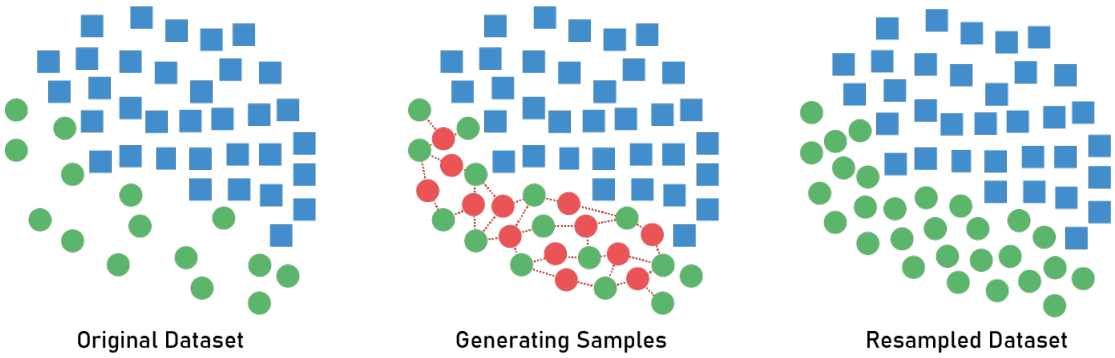


Figure 4.4: SMOTE Example

4.5.2 Validity Testing

After applying SMOTE, we used Logistical Regression and Random Forest Classification to analyze the augmented data and check for accuracy to ensure that the balanced dataset is usable to find correct prediction and correlation patterns in the features.
The Logistical Regression testing showed from 35% to 50% improvement in accuracy while the Random Forest Classification produced accuracy score improvement from 35.6% to 75%.

In figure 4.5 we can see the accuracy of the data from Random Forest Classifier and Logistic Regression as a graphical representation.

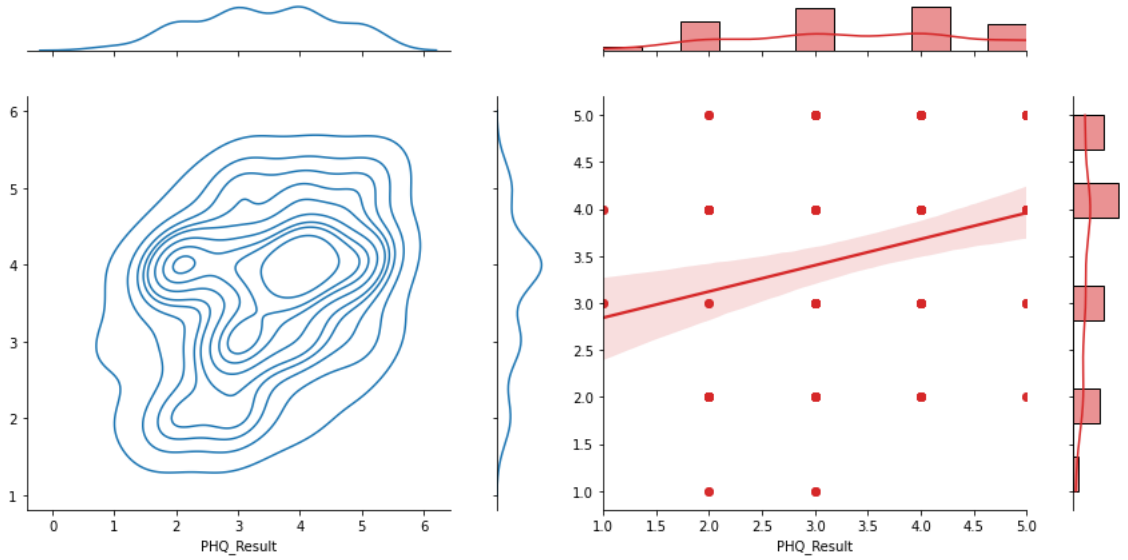


Figure 4.5: Pre Augmentation accuracy using Random Forest Classifier and Logistic Regression

Now, after the augmentation, the data accuracy increases.

In figure 4.6 we can see graphical representation for the accuracy result of augmented data estimated by Random Forest Classifier and Logistic Regression.

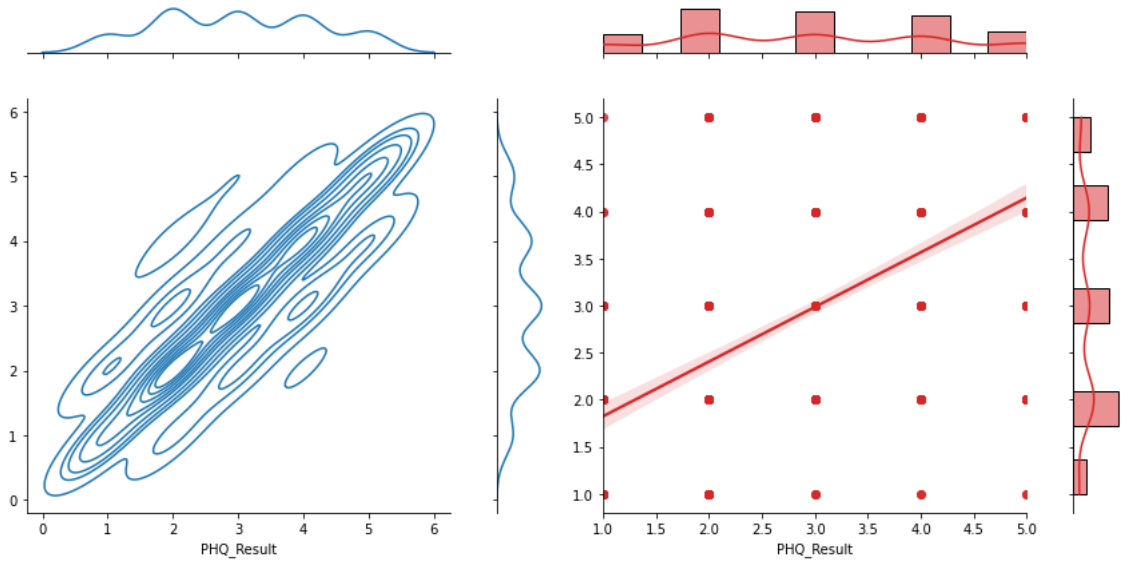


Figure 4.6: Post Augmentation accuracy using Random Forest Classifier and Logistic Regression

This is due to the fact that, initially the dataset was imbalanced and after augmentation and implementing SMOTE, the algorithms received a more balanced data to process. As for augmentation features, SMOTE was performed on the minority classes of Rural area and data collected from university outside of Dhaka. We can safely make assumption that for a larger dataset the algorithm model will produce even more accurate results.

4.6 Feature Selection

While creating a predictive model, the number of input variables should be reduced to lower the computational cost of modeling and, in some situations, to increase the model's performance. The process of minimizing the number of input variables is feature selection.
As the dataset is labeled, feature engineering is not needed. Only feature selection is necessary to negate some less effective features.
After pre-processing we obtain a dataset of (2108 , 28). Which means there are 27 features excluding the 'PHQ_Result'.

As the dataset has too many features, heat map in figure 4.7 is difficult to interpret. There are many relations that have values from 0.0 to 0.4, which are weakly correlated to each other. To make it simpler,distribution plots were used initially to neglect some features that have lighter density.

Distribution plots compare the empirical distribution of sample data with the theoretical values anticipated from a specific distribution to visually analyze the distribution of sample data. To assess whether the sample data originates from a specific distribution, we use distribution charts, in addition to more formal hypothesis testing.

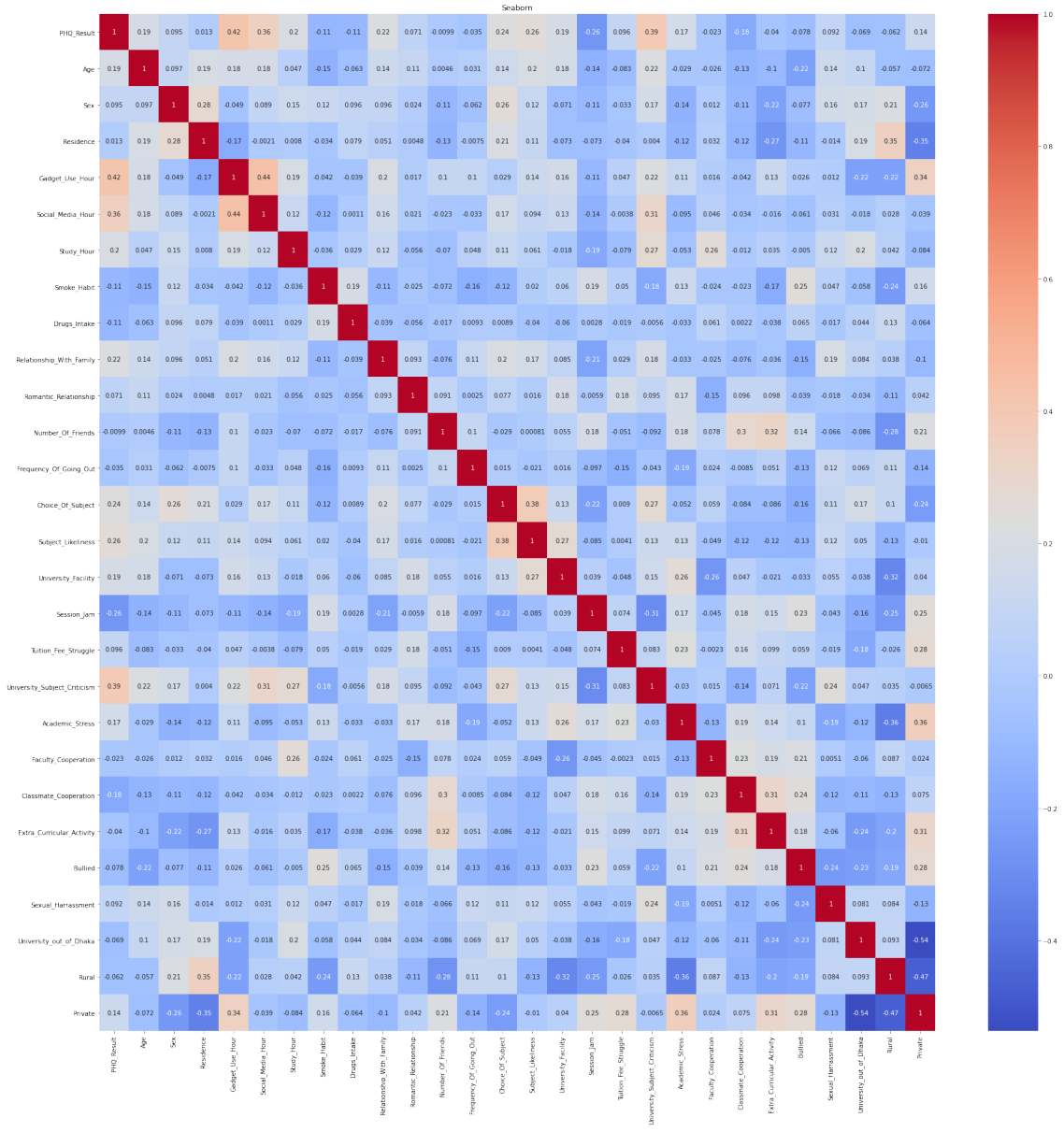


Figure 4.7: Heat Map of the data set before feature selection

In the data set, 'PHQ_Result' is the target and the remaining 27 columns are labels. We used *sns.jointplot()* to compare the distribution of each label against the distribution of the target, where hue = 'Private' (1= private, 0= public). Distribution plotting is useful to find out which labels affect the distribution the least. This information can be applied to eliminate some unimportant features and make the data cleaner.

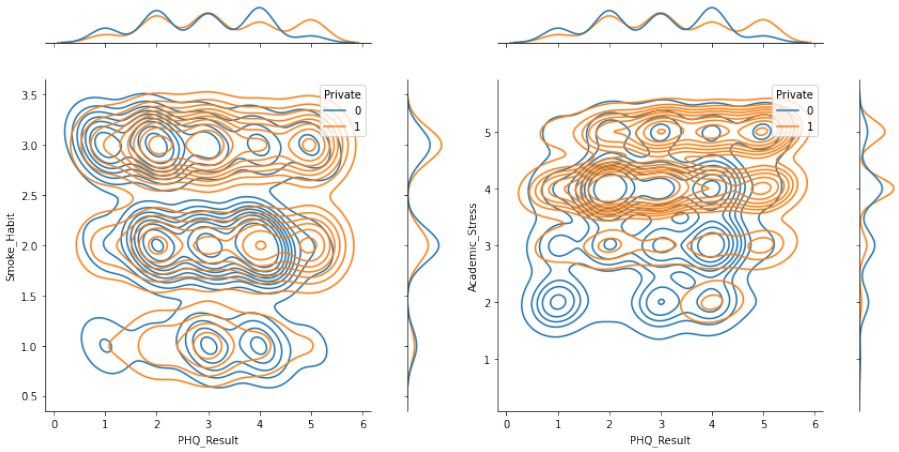


Figure 4.8: Strongly correlated features with PHQ-9 Result

On the other hand, some distribution plot that have lower weight are shown in figure 4.9:

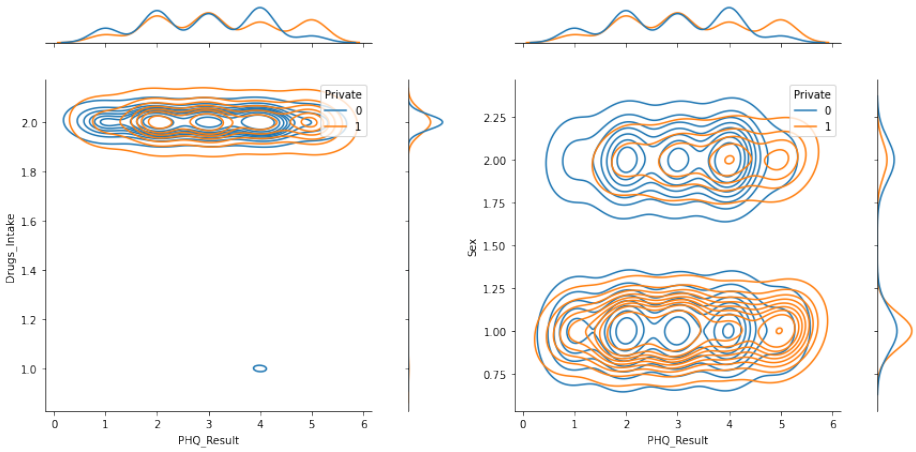


Figure 4.9: Weakly correlated features with PHQ-9 Result

Feature selection was done based on the result of Distribution plotting.
After dropping the less important labels, the features that were finalized are in figure 4.10.
After feature selection and re-plotting, the heat map in figure 4.11 looks cleaner and easier to describe, as the features present here have high correlation value.

```
Index(['Residence', 'Gadget_Use_Hour', 'Social_Media_Hour', 'Study_Hour',
       'Smoke_Habit', 'Romantic_Relationship', 'Number_Of_Friends',
       'Frequency_Of_Going_Out', 'Choice_Of_Subject', 'Session_Jam',
       'Tuition_Fee_Struggle', 'Academic_Stress', 'Faculty_Cooperation',
       'Extra_Curricular_Activity', 'Bullied', 'University_out_of_Dhaka',
       'Private'],
      dtype='object')
```

Figure 4.10: Selected Features

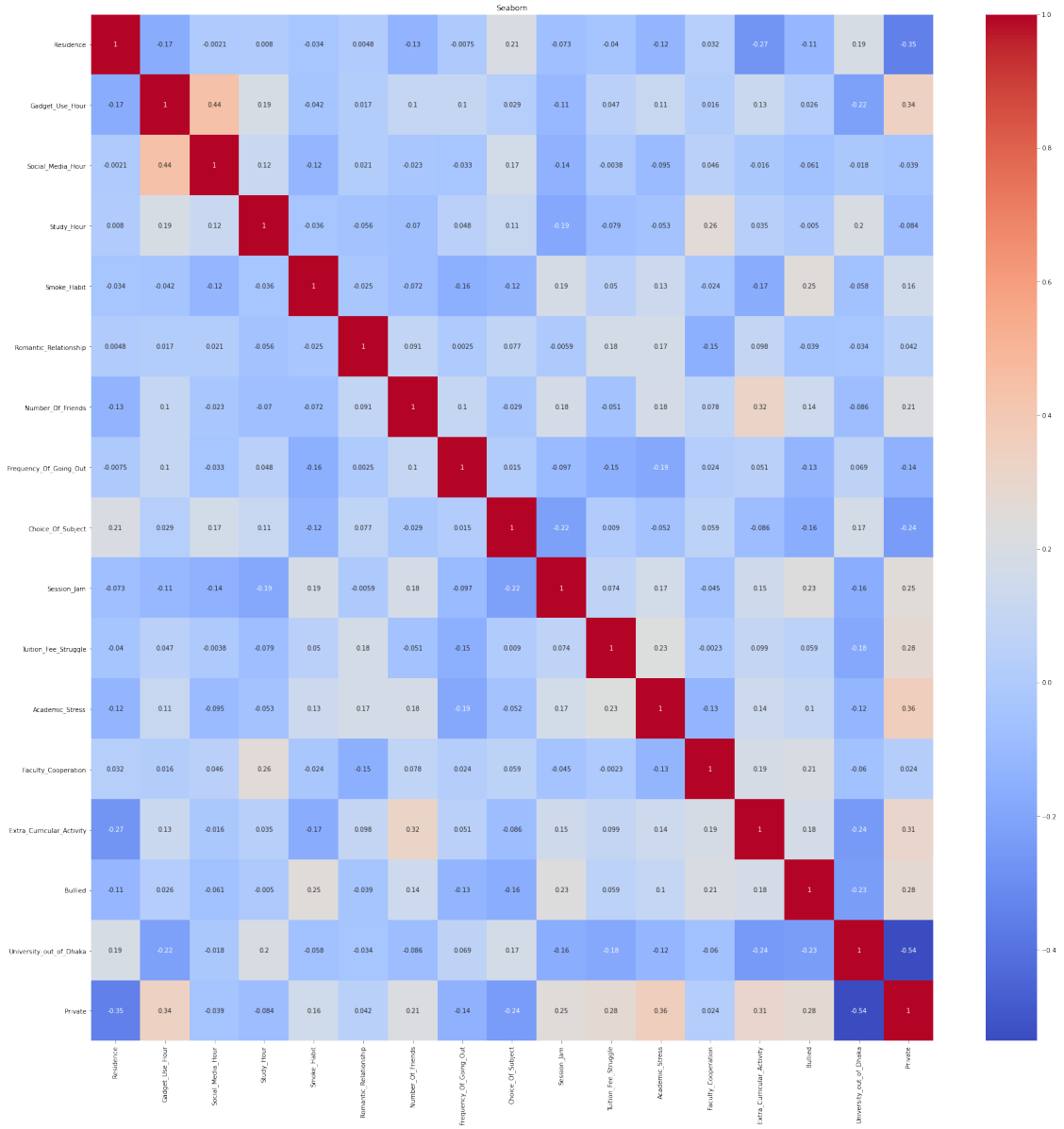


Figure 4.11: Heat Map of the data set after feature selection

4.7 Public Private Splitting

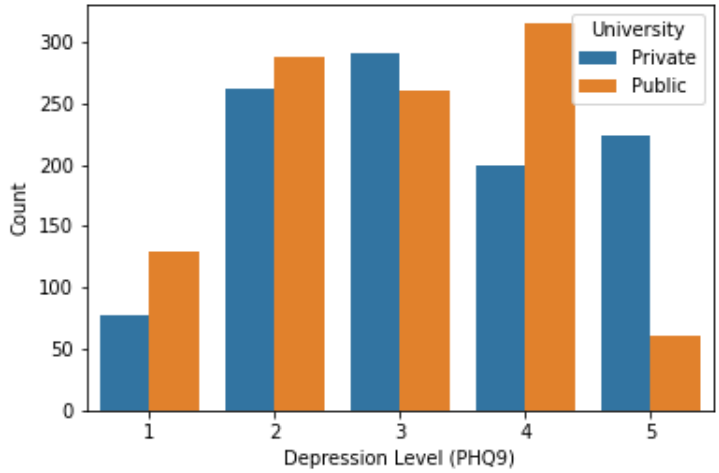


Figure 4.12: Count Plot demonstrating PHQ-9 dataset with separated Public and Private university data

After ensuring that the dataset is valid and we have a reasonable size of data to fit into the model, the dataset was split into two parts according to the "public_or_private" column of the dataset. In this research, we fit the dataset into different machine learning models to evaluate the models' accuracy and error. After evaluation, during the final implementation and result analysis part, the large dataset was split into two segments, as per the public and private university attribute.

4.8 Test and train split

After feature selection the dataset was split into train and test set to conduct model evaluation. 'PHQ_Result' is the target column (y) and the rest of the columns are labels(X). The X_train, y_train sets are used for training the models. And, to test the performance of the models, X_test, y_test are used. The splitting is done in 80:20 train-test ratio. Meaning 80% of the data is being used for training purposes while the rest 20% is kept for testing the MAE and R2 value of the algorithm. After splitting the data set, it gives a Baseline Performance of 0.925 Mean Absolute Error.

Chapter 5

Model Implementation & Optimization

5.1 Work Flow Overview

To build a supervised model that can find out the important features and identify the associations with the target, we need to select a regression algorithm and a mining algorithm. Not to mention, selecting the right algorithms is very important to do so. After pre-processing and cleaning the initial dataset, we have fitted the data in a few regression models to evaluate their performances. The one with the least error and most accuracy has been selected to conduct the research. As for the mining part we have selected Apriori Algorithm as it is the most suited to apply on our dataset.
Here is a detailed description of the workflow of our research:

- Firstly the collected data set is label encoded. We convert the data from text to numerical and encode the labels to our convenience.
- Next we use Synthetic Minority Oversampling Technique (SMOTE) to resolve any sort of imbalance in the dataset that might cause the model to perform poorly.
- PHQ-9 algorithm is applied in the later stage. The first nine columns that contain the user inputs of the PHQ-9 questionnaire, are taken into consideration. After getting the results, the mentioned columns are replaced with 'PHQ_Result'
- Next one is Test-train split. As the name suggests, the entire data is split into training sets and testing sets to evaluate the performances of the models when the data is fitted into them.
- Model evaluation is an essential part where we train some models and analyze their performances by detecting their Mean Absolute error and R2. In this step we select the correct regression algorithm that can work best for our research.
- Next, the entire dataset is equally split into public university data and private university data to fit them in the regression part of our model. But to use the data for mining purposes, the entire dataset is 1-Hot Encoded before splitting it into public and private.

- In the following stage, we find out the important features from the regression process, and then mine the data using Apriori Algorithm to find the frequently occurring factors and show their association with the target.
- After that, we analyze the results of data mining and identify strongly correlated factors that have prominent effect and influence on depression level within both datasets
- Lastly, we do a comparative discussion on the results from both university data and draw a finish line by mentioning the problems that university students are facing which are disrupting their mental health.

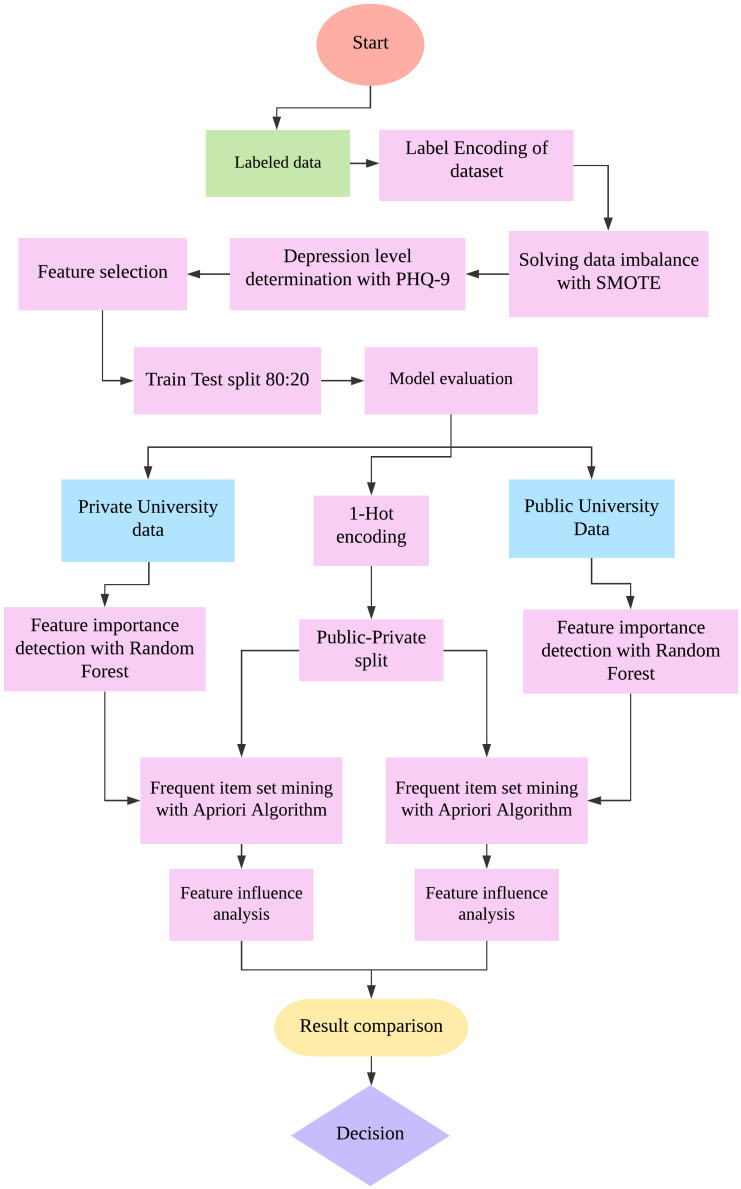


Figure 5.1: Workflow

5.2 Model Evaluation

For the regression task in our supervised model we need to find the best regression algorithm that can do the work more promisingly. After splitting the data into test and train sets, the baseline error was 0.925 based on the Mean Absolute Error. In this part of the research, we build three models with three regression algorithms and observe their performances.

1. Logistic Regression
2. Random Forest Regression
3. K-Nearest Neighbors Regression

Using the Scikit-Learn library, we created three models with the stated algorithms and detected the Mean Absolute Error (MAE) and R2 for each of the models.

After fitting our dataset into three regression algorithms, we observe error values as shown in table 5.1.

Model	MAE	R2
Logistic Regression	0.688	0.226
Random Forest Regression	0.429	0.693
K-Nearest Neighbor	0.458	0.671

Table 5.1: Model Performance evaluation for testing set

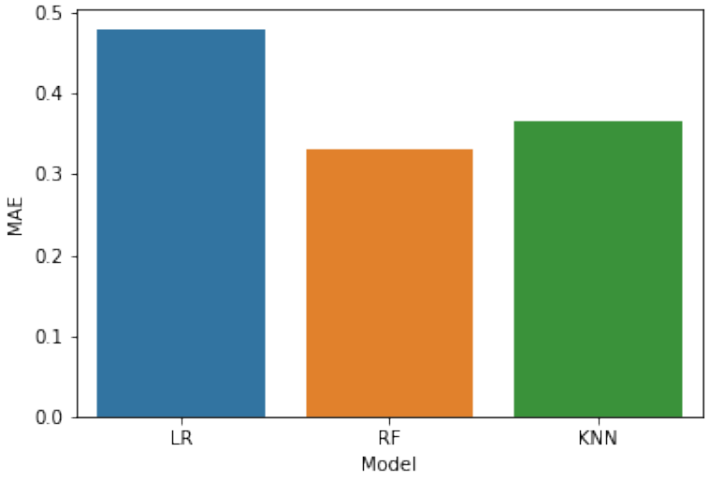


Figure 5.2: Mean Absolute Error evaluation for three models

Aforementioned calculated data and figure 5.2 shows that Logistic Regression gives the highest MAE 0.6884 whereas Random Forest Regressor generates the lowest MAE 0.434. K-nearest Neighbour scored 0.4580.

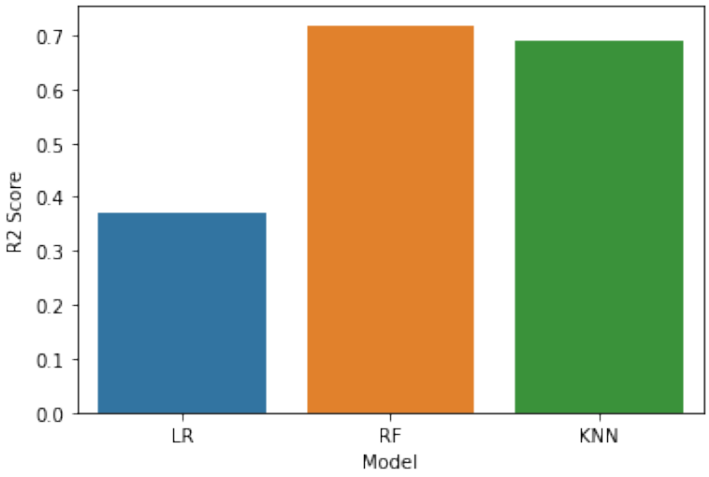


Figure 5.3: R2 evaluation for three models

Furthermore, figure 5.3 shows that the highest scoring algorithm is Random Forest Regression with the score of 0.6914 R2. And the least scoring model is Linear Regression with 0.2260 R2.

From these comparisons we came to a conclusion that Random Forest Regression has the most promising performance. It has the lowest MAE value. Moreover this algorithm scored the highest R2 value. On the other hand Logistic Regression and K-nearest Neighbor Regression did not reach the benchmark to be used in the regression part of our supervised learning model, in comparison with Random Forest Regression.

5.3 Proposed Hybrid Model

The proposed model consists of one regression algorithm and a mining algorithm as described in Project Workflow. As per the model evaluation process, the most suitable regression learning system to apply in our dataset is Random Forest. Figure 5.4 demonstrates the regression process, for which Random Forest Regression has been used, and the data mining process, which is done by the Apriori algorithm.

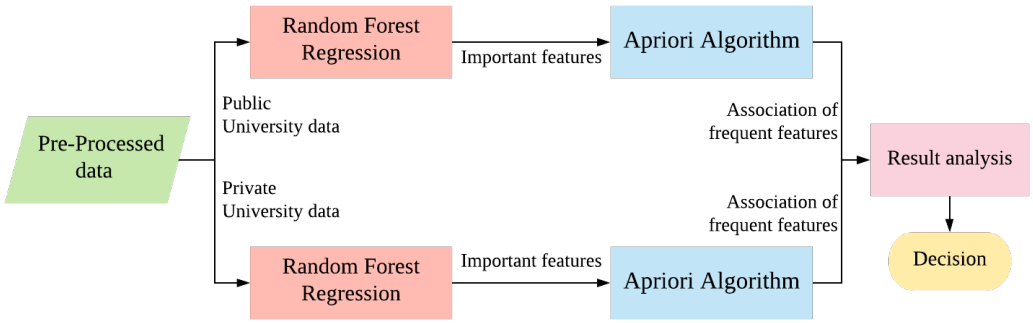


Figure 5.4: Propsed Hybrid Model

5.3.1 Random Forest Implementation

At first, to find out the important features that have the most influence on the depression levels ('PHQ_Result'), the data set is split into two individual datasets, public university data and private university data. Each of the dataset is then train-test split to fit in the Random Forest Regression model.

In the dataset we divided the train sets and test sets into 80:20 ratios for both datasets, public and private data. While fitting the train sets into the model, we tuned the number of decision trees and random states to acquire maximum accuracy. The least Mean Absolute Error is gained when the algorithm generates 1000 decision trees with 0 random states. Figures 5.5 and 5.6 are two Decision Tree visualizations from each data sets within max_depth = 3 and random_state = 100

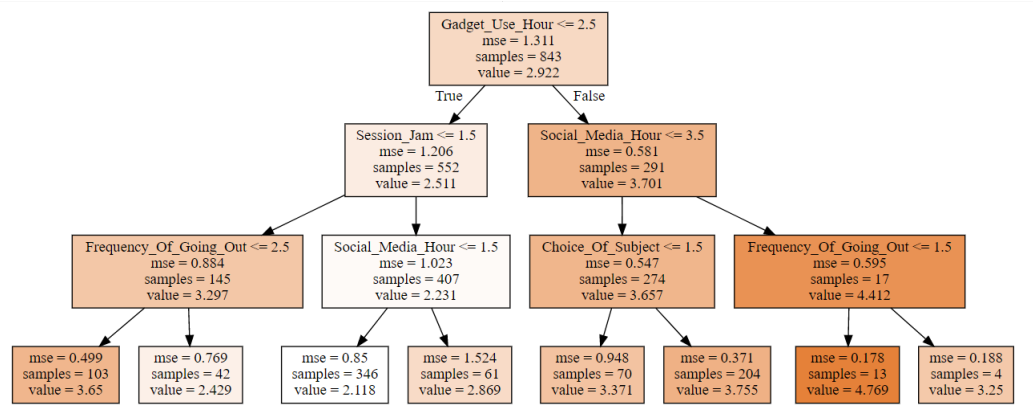


Figure 5.5: Single Decision Tree from public university data

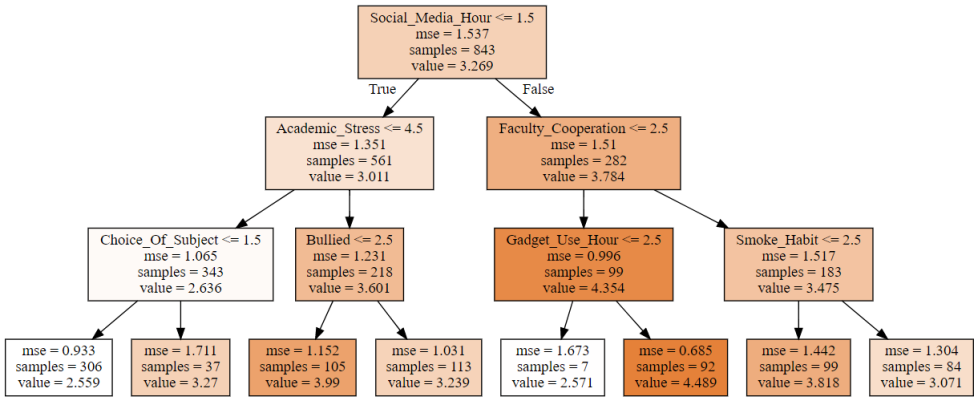


Figure 5.6: Single Decision Tree from private university data

In the new step of the model we found the most important features that affect the 'PHQ_Result'. Tables 5.2 and 5.3 are two lists of those features in both data sets:

Variable	**Importance**
Gadget_Use_Hour	0.27
Faculty_Cooperation	0.11
Session_Jam	0.09
Academic_Stress	0.09
Number_Of_Friends	0.05
Choice_Of_Subject	0.05
Extra_Curricular_Activity	0.05
Social_Media_Hour	0.04
Smoke_Habit	0.04
Frequency_Of_Going_Out	0.04
Tuition_Fee_Struggle	0.04
Bullied	0.04
Romantic_Relationship	0.03
Residence	0.02
Study_Hour	0.02
University_Location	0.02

Table 5.2: Variable Importance list of the Public University data

Variable	**Importance**
Academic_Stress	0.12
Faculty_Cooperation	0.11
Extra_Curricular_Activity	0.10
Social_Media_Hour	0.09
Gadget_Use_Hour	0.08
Number_Of_Friends	0.07
Bullied	0.07
Romantic_Relationship	0.06
Choice_Of_Subject	0.06
Tuition_Fee_Struggle	0.06
Smoke_Habit	0.05
Frequency_Of_Going_Out	0.04
Study_Hour	0.03
Session_Jam	0.02
University_Location	0.02
Residence	0.01

Table 5.3: Variable Importance list of the Private University data

Following this, the model is again trained with the most important features for better results. This time the model gives more accuracy. The new Mean Absolute Error: 0.94 degrees and Accuracy: 72.09 %.

The final visualization of the results are demonstrated in figure 5.7

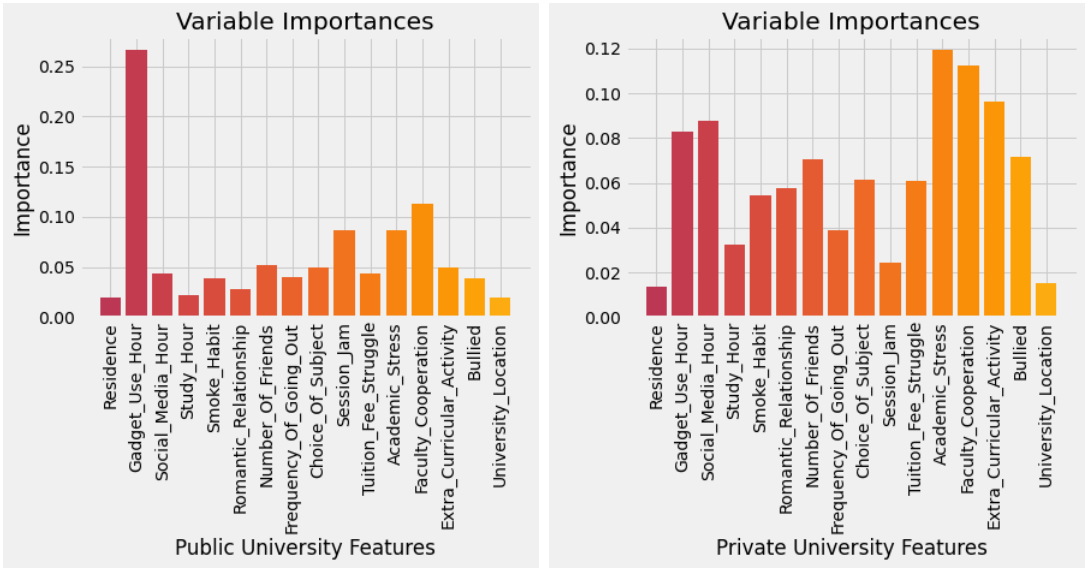


Figure 5.7: Variable importance [Visualized]

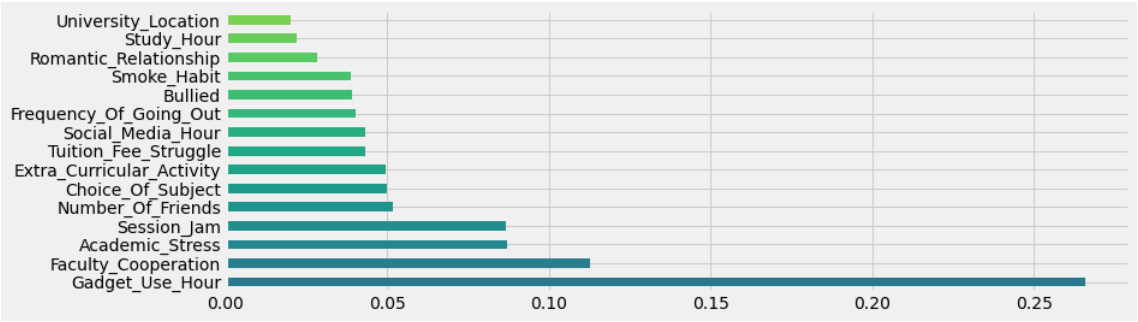


Figure 5.8: Variable importance [Public] [Visualized, Sorted]

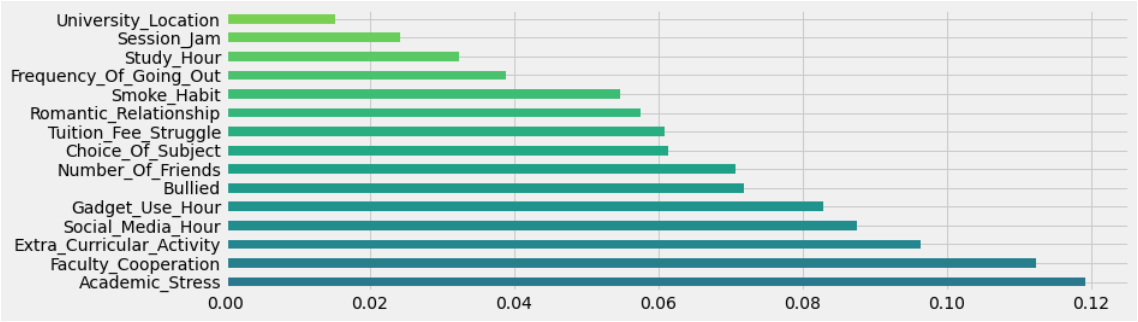


Figure 5.9: Variable importance [Private] [Visualized, Sorted]

Figures 5.8 and 5.9 assist to get an easier look on the result. As seen from Public University variable importance, the most prominent variable that affects the mental condition of public university students is 'Gadget_Use_Hour' which has an importance of 0.27. Even though Faculty_cooperation is in second place, the difference among the importance of these top two variables is very significant. Not only that, Gadget_use_hour of public university data even surpasses the top important variable of private university data, 'Academic_Stress', that has an importance of 0.12.
On the other hand, in figure 5.9 we can see a more closely related chronological order among the private university variables. The differences among their values are not prominent.

5.3.2 Limitation of Random Forest

Random Forest is an effective solution to bring up the prominent features in the data. But figuring out only the label importance is not enough to reach the goal of this research. To get a better understanding of the scenario, we need to find the association among the variables and the mental state of the students. Which is why random forests can not draw the end line of this research. To further clear it out, in order to find associations, we also have to consider the parameters of each label. For instance, we now understand that 'Number_Of_Friends' is moderately important for both the universities. But the connection with the parameters of this feature : No friend, 2-3 friends, Group of 5-7 friends, Multiple friend groups can not be generated with random forest. But to understand why a university student may suffer from mental turbulence, it is essential to pinpoint the exact attributes of the personal and university lives of the students.

5.3.3 Frequent Features Mining

As per our previous findings, we have found the important factors among our dataset that are strongly correlated to the depression levels in both public university students and private university students. However, the findings do not completely foretell the level of influence for those factors, as each factor has a positive and negative weighted value, our research also aims to find the negative factors that draw influence over the students. We can use another algorithm that is used for frequent itemsets mining and estimating the associations among features in a dataset, to solidify our predicted factor influence in depression level.

5.3.4 Apriori Algorithm Implementation

Apriori algorithm is a popular approach that analyzes the associations among the features in large datasets and tries to establish a numerical value to these associations as per the frequency and correlation of each feature's appearance in the dataset. Apriori uses the "market basket" analysis in data mining to find frequent itemsets from the estimated association rules and tries to predict a generic occurrence pattern. We are using Apriori algorithm to find the association among the important factors in our dataset of both public and private university students and then filter out the association between different PHQ-9 based data and other factors. Basically

we try to determine the associations between five levels of PHQ-9-estimated depression level in our data and whichever factors are most frequent and predicatively most influential. As discussed earlier, Apriori Algorithm is a data mining approach that finds the frequent itemsets using 'market-basket' method and also estimates 'support', 'confidence' and 'lift' between two items.

5.3.5 Association Rule & Frequent itemset mining

Association rule refers to the correlation between two itemsets in a given data. If there is an association rule of $A -> B$, it signifies that there is a likelihood of B occurring whenever A is present. 'Confidence' is a value that indicates the likelihood of a particular item appearing when a former item appears. Apriori Algorithm estimates all the rules into a frozen-itemset list. Then Apriori finds the support of each item and confidence of every item's correlation with other items. In this research, we are using Apriori algorithm to generate a list of association rules between every columns in our dataset. Running with a minimum support of 0.05 and a minimum confidence threshold of 0.3, we generate a large data of association rules.

```
[169] from mlxtend.frequent_patterns import association_rules
      res = association_rules(fritems, metric='confidence', min_threshold=0.3)
```

```
[170] res
```

	antecedents	consequents
0	(PHQ_Minimal_Depression)	(Self_Chosen_Subject)
1	(PHQ_Minimal_Depression)	(Social_Media_Usage_0-5)
2	(PHQ_Minimal_Depression)	(Study_Hour_0-5)
3	(PHQ_Minimal_Depression)	(High_Academic_Stress)
4	(PHQ_Minimal_Depression)	(Urban)
...	...	...
353956	(Low_Extra_Curricular_Involvement, Rare_Social...	(Study_Hour_0-5, Non_Smoker, Urban, High_Acade...
353957	(Self_Chosen_Subject, Low_Extra_Curricular_Inv...	(Study_Hour_0-5, Non_Smoker, Urban, High_Acade...
353958	(Self_Chosen_Subject, Low_Extra_Curricular_Inv...	(Study_Hour_0-5, Non_Smoker, Urban, High_Acade...
353959	(Low_Extra_Curricular_Involvement, Non_Smoker,...	(Study_Hour_0-5, Urban, High_Academic_Stress, ...
353960	(Low_Extra_Curricular_Involvement, Rare_Social...	(Study_Hour_0-5, Non_Smoker, Urban, High_Acade...

353961 rows × 9 columns

Figure 5.10: Association Rules data-mined by Apriori Algorithm

In the figure 5.10, we find a large association rule data of frozen itemsets has been generated. For further inspection, we filter association rules where five levels of PHQ-9 values are in 'antecedents', itemset length is 2 and find that factors with importance value of 0.4 or higher have the most frequency and occurrence in Apriori algorithm result. Academic Stress, Subject Choice, Bullying, Number of Friends, Gadget Usage, Social Media Usage, Extra-curricular Activity, Faculty Cooperation, Smoking, Session Jam - these ten factors are finalized for result analysis.

Chapter 6

Result Analysis

In chapter 5, we estimated importance of labeled factors in our dataset through Random Forest Regression, and high importance features were selected. Using Apriori, we then found association rules between five separate depression levels and the aforementioned selected features. After association rules mining, our model found that severe_depression levels had no association above the minimum support or confidence threshold. Therefore, we have no association rules for severe depression in public universities.

6.1 Association Rules Mining for Public University Students Data

In the tables 6.1, 6.2, 6.3 and 6.4 we can see the support, confidence and lift values of selected factors against the range of depression levels in our public university dataset.

	Support	Confidence	Lift
PHQ_Minimal_Depression			
Gadget_Usage_0-5	0.055977	0.453846	1.523420
Gadget_Usage_5-10	0.065465	0.530769	1.491815
Moderate_Cooperative_Faculties	0.054080	0.438462	1.015689
Minimal_Friends	0.070209	0.569231	1.199938
Self_Chosen_Subject	0.111954	0.907692	2.116610
Low_Extra_Curricular_Involvement	0.085389	0.692308	1.002325
Bullied_as_as_Freshman	0.051233	0.415385	1.426109
Non_Smoker	0.089184	0.723077	2.021547

Table 6.1: Association itemsets for PHQ_Minimal_Depression in Public dataset

	Support	Confidence	Lift
PHQ_Mild_Depression			
Gadget_Usage_0-5	0.148956	0.545139	1.829861
Gadget_Usage_5-10	0.102467	0.375000	1.054000
Less_Cooperative_Faculties	0.115750	0.423611	1.313194
Moderate_Cooperative_Faculties	0.145161	0.531250	1.230632
High_Academic_Stress	0.207780	0.760417	1.381861
Minimal_Friends	0.130930	0.479167	1.010083
Self_Chosen_Subject	0.148008	0.541667	1.263090
Subject_by_Merit_Position	0.123340	0.451389	0.828857
Low_Extra_Curricular_Involvement	0.211575	0.774306	1.121041
Bullied_as_as_Freshman	0.090133	0.329861	1.132487
Irregular_Smoker	0.123340	0.451389	0.913174
Non_Smoker	0.137571	0.503472	1.407585

Table 6.2: Association itemsets for PHQ_Mild_Depression in Public dataset

	Support	Confidence	Lift
PHQ_Moderate_Depression			
Gadget_Usage_10-15	0.090133	0.365385	1.356040
Moderate_Cooperative_Faculties	0.125237	0.507692	1.176061
High_Academic_Stress	0.128083	0.519231	0.943568
Faced_Session_Jam	0.126186	0.511538	1.404067
Minimal_Friends	0.134725	0.546154	1.151292
Self_Chosen_Subject	0.077799	0.315385	0.735432
Subject_by_Merit_Position	0.161290	0.653846	1.200616
Low_Extra_Curricular_Involvement	0.152751	0.619231	0.896524
Bullied_as_as_Freshman	0.077799	0.315385	1.082786
Bullied_Often	0.124288	0.503846	1.341045
Irregular_Smoker	0.135674	0.550000	1.112668

Table 6.3: Association itemsets for PHQ_Moderate_Depression in Public dataset

	Support	Confidence	Lift
PHQ_Moderately_Severe_Depression			
Gadget_Usage_5-10	0.101518	0.339683	0.954734
Gadget_Usage_10-15	0.134725	0.450794	1.673016
Moderate_Academic_Stress	0.120493	0.403175	1.470401
High_Academic_Stress	0.133776	0.447619	0.813432
Faced_Session_Jam	0.193548	0.647619	1.777579
Minimal_Friends	0.122391	0.409524	0.863276
Group_Of_Friends	0.124288	0.415873	1.138520
Subject_by_Merit_Position	0.223909	0.749206	1.375720
Low_Extra_Curricular_Involvement	0.202087	0.676190	0.978990
Bullied_Often	0.190702	0.638095	1.698365
Irregular_Smoker	0.203036	0.679365	1.374378

Table 6.4: Association itemsets for PHQ_Moderately_Severe_Depression in Public dataset

First of all, we see in figure 6.1 gadget usage shows a low downward fluctuation in the dataset, having max confidence around 0.5. This indicates that gadget_usage has no influence over public university students.

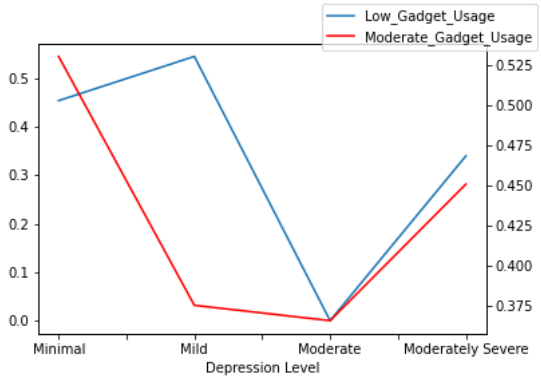


Figure 6.1: Gadget Usage in Public University

Next, we find in figure 6.2 that, faculty cooperation also demonstrates low downward slope, which states that faculty cooperation does not impact students' mental health in public universities.

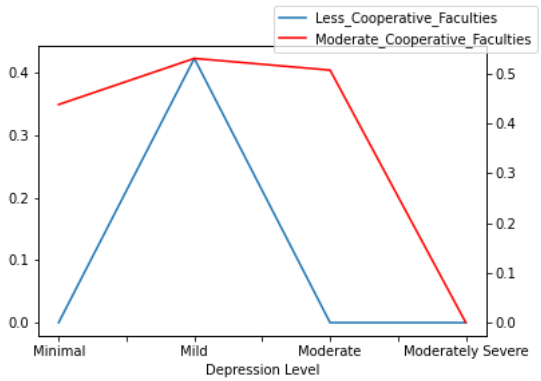


Figure 6.2: Faculty Cooperation in Public University

After that, we see academic stress having an upward line in figure 6.3, indicating that high depression level shows high confidence of academic stress association. So, academic stress is highly influential in public university data.

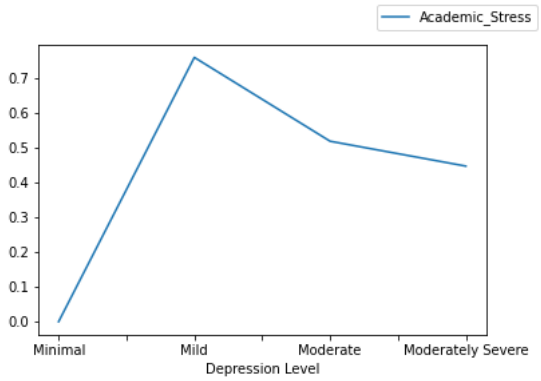


Figure 6.3: Academic Stress in Public University

Then, we see session jam confidence in figure 6.4, which exhibits an upward line graph, reaching above 0.6 confidence for moderately_severe depression. This indicates session jam having high influence on public university students' mental health.

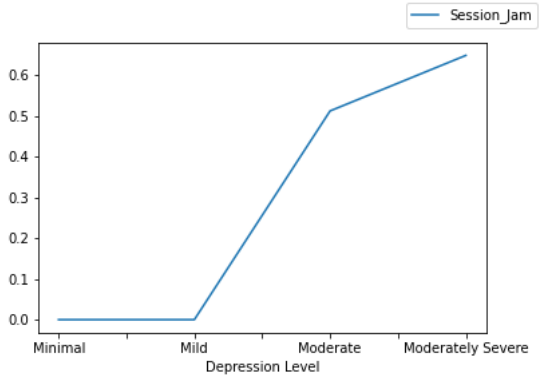


Figure 6.4: Session Jam in Public University

In figure 6.5, feature of "friends group" with minimal friends data shows a downward graph with less than 0.5 confidence throughout, proving that depression levels

in public universities are not strongly influenced by number of friends they have.

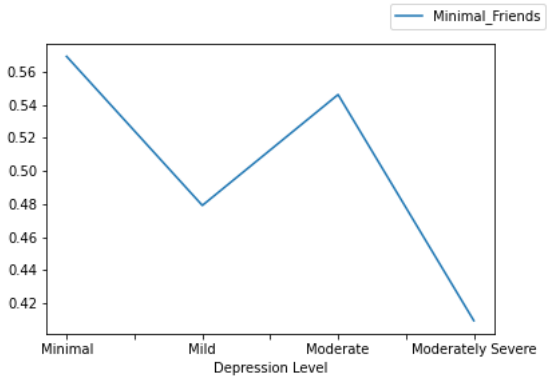


Figure 6.5: Friends Group in Public University

Furthermore, in figure 6.6, we observe two counter-reflective data for subject choice factors. Rise of depression level is parallel with an upward graph of merit based subject choice while self chosen subject choice displays opposite values. This indicates that merit based subject choice can often lead to high depression, while entertaining a student's own preference while choosing subject can result in low depression levels for public university students.

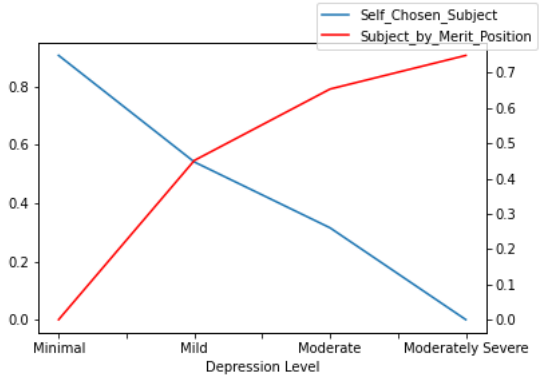


Figure 6.6: Subject Choice in Public University

After that, figure 6.7 shows a consistency of high confidence value throughout depression levels. This indicates that students of all depression levels have equally low extra-curricular involvement, therefore it has no influence on students.

Following, Figure 6.8 shows an increased association of bullying as depression levels rise. Naturally, in public universities, a culture of "ragging" persists that indirectly or directly appropriates bullying by seniors [29]. This form of bullying, as visualized, can and does lead to negatively influence a student's mental health.

Finally, in figure 6.9, we observe smoking-habit displaying parallel association with depression level while students that are non-smokers show an improvement in mental health.

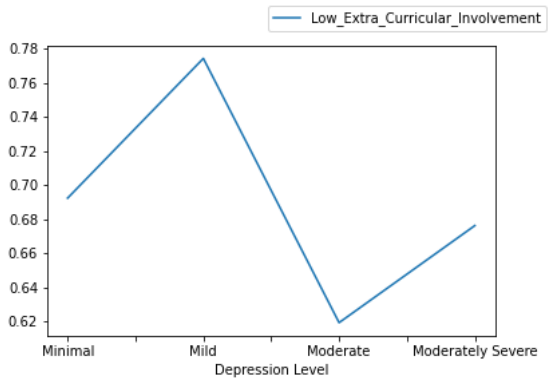


Figure 6.7: Extra Curricular Activity in Public University

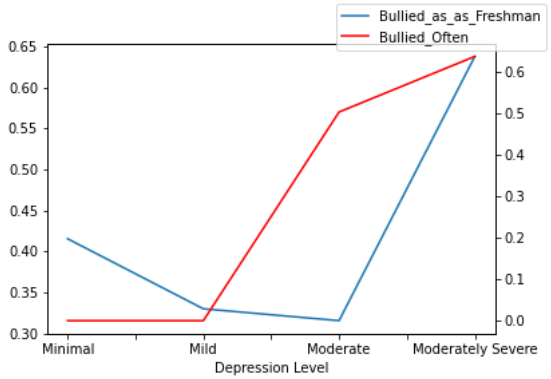


Figure 6.8: Bullying in Public University

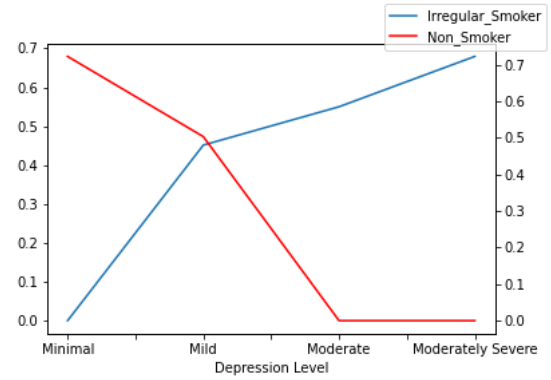


Figure 6.9: Smoking in Public University

6.2 Association Rules Mining for Private University Students Data

Similarly, we cross-referenced the factors analyzed in the random forest regression implementation with the Apriori algorithm's frequent itemset table and found that some factors with low importance score did not produce any association with PHQ-9 column data in our dataset. Thus, we tabulated the association rules that were above our confidence threshold of 0.5.

Below are tables 6.5, 6.6, 6.7, 6.8 and 6.9 that shows the data mined results, having

support, confidence and lift values between range of PHQ-9 depression levels and selected features in private university dataset.

	Support	Confidence	Lift
PHQ_Minimal_Depression			
High_Academic_Stress	0.055028	0.743590	0.866015
Social_Media_Usage_0-5	0.058824	0.794872	1.193440
Self_Chosen_Subject	0.066414	0.897436	1.119405

Table 6.5: Association itemsets for PHQ_Minimal_Depression in Private dataset

	Support	Confidence	Lift
PHQ_Mild_Depression			
High_Academic_Stress	0.215370	0.869732	1.012925
Moderate_Cooperative_Faculties	0.091082	0.367816	0.893268
Social_Media_Usage_0-5	0.194497	0.785441	1.179280
Gadget_Usage_5-10	0.111954	0.452107	1.498494
Gadget_Usage_10-15	0.094877	0.383142	0.746454
Group_Of_Friends	0.104364	0.421456	1.105011
Self_Chosen_Subject	0.223909	0.904215	1.127861
Irregular_Smoker	0.075901	0.306513	0.915199
Non_Smoker	0.144213	0.582375	1.063819

Table 6.6: Association itemsets for PHQ_Mild_Depression in Private dataset

	Support	Confidence	Lift
PHQ_Moderate_Depression			
High_Academic_Stress	0.256167	0.927835	1.080595
Moderate_Cooperative_Faculties	0.142315	0.515464	1.251841
Social_Media_Usage_0-5	0.214421	0.776632	1.166055
Gadget_Usage_10-15	0.152751	0.553265	1.077894
Group_Of_Friends	0.114801	0.415808	1.090202
Self_Chosen_Subject	0.233397	0.845361	1.054450
Non_Smoker	0.166983	0.604811	1.104802

Table 6.7: Association itemsets for PHQ_Moderate_Depression in Private dataset

	Support	Confidence	Lift
PHQ_Moderately_Severe_Depression			
High_Academic_Stress	0.148956	0.785	0.914243
Less_Cooperative_Faculties	0.062619	0.330	1.159400
Moderate_Cooperative_Faculties	0.073055	0.385	0.935000
Social_Media_Usage_0-5	0.110057	0.580	0.870826
Social_Media_Usage_5-10	0.062619	0.330	1.312528
Gadget_Usage_10-15	0.090133	0.475	0.925416
Minimal_Friends	0.074004	0.390	1.347738
Group_Of_Friends	0.061670	0.325	0.852114
Self_Chosen_Subject	0.133776	0.705	0.879373
Irregular_Smoker	0.085389	0.450	1.343626
Non_Smoker	0.082543	0.435	0.794610

Table 6.8: Association itemsets for PHQ_Moderately_Severe_Depression in Private dataset

	Support	Confidence	Lift
PHQ_Severe_Depression			
High_Academic_Stress	0.183112	0.861607	1.003463
Less_Cooperative_Faculties	0.096774	0.455357	1.599821
Moderate_Cooperative_Faculties	0.080645	0.379464	0.921556
Social_Media_Usage_0-5	0.088235	0.415179	0.623359
Social_Media_Usage_5-10	0.075901	0.357143	1.420485
Gadget_Usage_10-15	0.141366	0.665179	1.295930
Bullied_Often	0.075901	0.357143	1.726737
Minimal_Friends	0.080645	0.379464	1.311329
Self_Chosen_Subject	0.144213	0.678571	0.846407
Irregular_Smoker	0.091082	0.428571	1.279644
Non_Smoker	0.112903	0.531250	0.970429

Table 6.9: Association itemsets for PHQ_Severe_Depression in Private dataset

Moving onto association analysis of private university data, we can firstly observe in figure 6.10, an alarming high confidence value for academic stress (between 70% to 90%), which parallels depression level. This marks a high influence of academic stress in private university students' mental health.

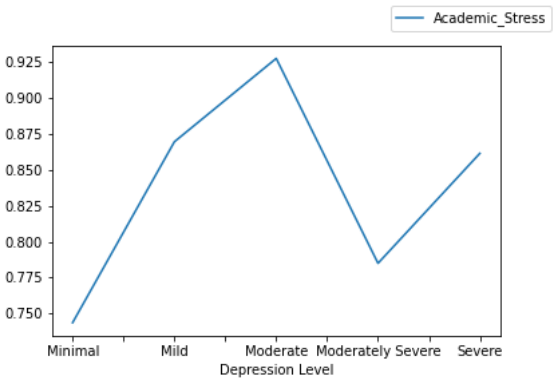


Figure 6.10: Academic Stress in Private University

Secondly, in figure 6.11, faculty cooperation has a low confidence value, but the graph displays an upward line-graph. This states a moderately associated influence of faculty-cooperation on student's mental health in private universities.

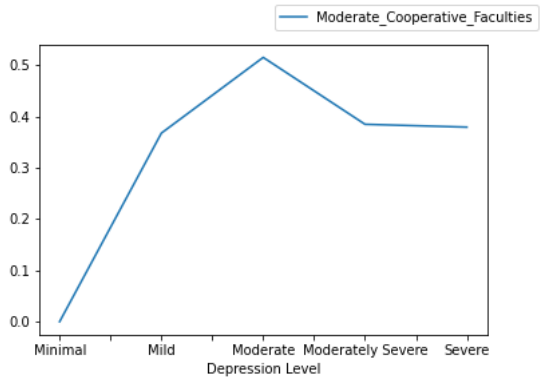


Figure 6.11: Faculty Cooperation in Private University

Thirdly, social media usage statistics in figure 6.12 exhibits a strong association between social media use and depression level. Rising depression levels from minimal to severe is linked with an increase in social media usage, as observed.

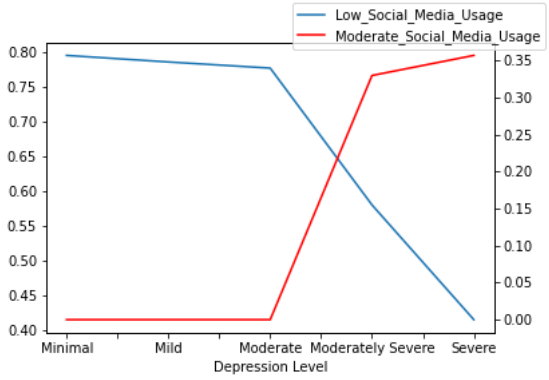


Figure 6.12: Social Media Usage in Private University

Similarly, we observe an upward graph in figure 6.13, showing parallel relation between moderate gadget usage and depression level. Hence, gadget usage is also

influential on private university student's mental turbulence.

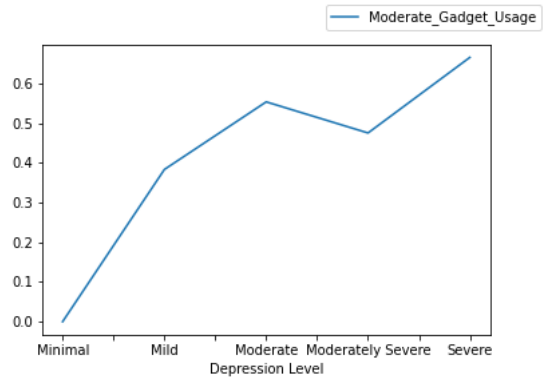


Figure 6.13: Gadget Usage in Private University

After that, in figure 6.14, friends group size reflects on a negative impact on students' mental health. Graph demonstrates number of friends significantly decreasing when the depression level increases.

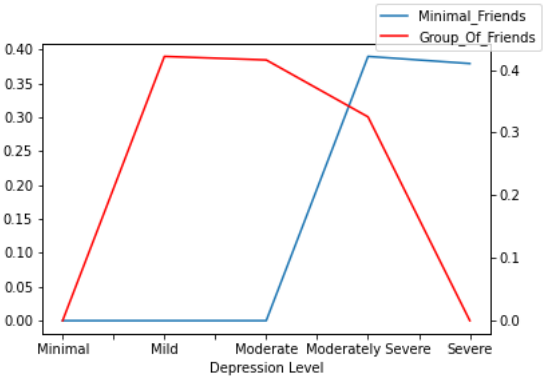


Figure 6.14: Friends Group in Private University

Next, we discuss the subject choice feature in private universities. As most private universities allow students to choose their subjects and departments with minimum criteria, figure 6.15 has only self chosen subject in the line graph. Also, it indicates a downward line-graph which states that even in high associative factors, subject choice draws no influence on depression level severity.

Finally, figure 6.16 displays a low confidence value graph for smoking habit, and both smoking and non-smoking displays similar upward rise. This indicates that both smoking and non-smoking are low-associated with depression level, proving no influence of such feature in private university student's mental health.

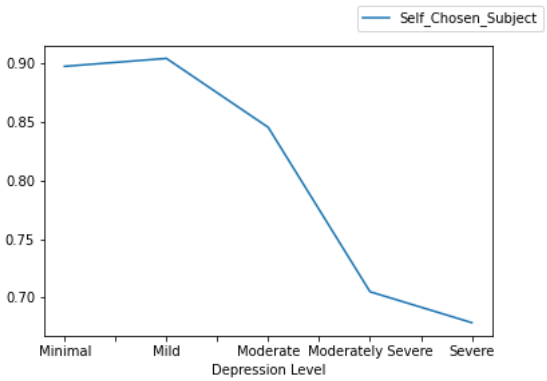


Figure 6.15: Subject Choice in Private University

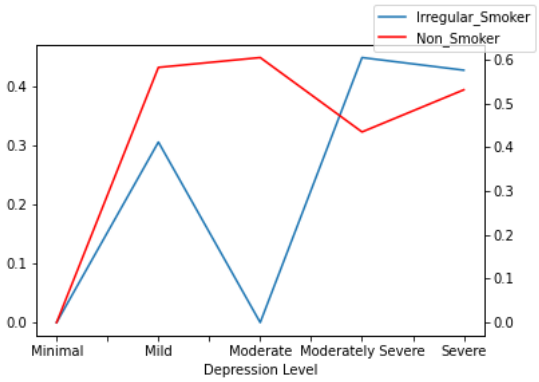


Figure 6.16: Smoking in Private University

6.3 Comparative Result Analysis

Finally, we have in table 6.10 a comparative analysis of influential factors in both public and private university datasets.

Feature	**Influence on Public University Students**	**Influence on Private University Students**
Academic Stress	High	High
Subject Choice	Negative (Merit-Position)	Positive (Self-Chosen)
Bullying	High	Low
Number of Friends	No	Moderate
Gadget Usage	No	Moderate
Social Media Usage	No	High
Extra-curricular Activity	No	No
Faculty Cooperation	No	Moderate
Smoking	High	No
Session Jam	High	No

Table 6.10: Comparative Analysis of features' influence in different university students

As per the comparative table, we observe merit based subject choice system, bul-

lying, smoking and session-jam factors have high influence on Public University students' mental health. On the other hand, number of friends, faculty cooperation, gadget usage and social media usage has moderate to high level of influence on private university's student mental well-being. Academic stress has high influence on both type of universities, indicating an extreme educational environment for students all around the country.

Additional discussion can be made on subject choice preference, which shows contrasting influence. Where public universities have restricted subjects accessible only through merit positions, students face frustration and pressure during admission exam session, leading to suicides and depression[19], as well as studying in subjects outside of their preference leading to mental struggle [12]. On the contrary, all-accessible subjects and departments allow private university students to receive credit transfer facilities and independent educative environment[2].

Chapter 7

Conclusion and Future Work

7.1 Conclusion

As the age of modern technology advances, we gradually realize the importance of mental health in the prosperity of civilization and development of human characteristics. Being a member of social construct, mental health issues are a primary factor in someone's behavior, interaction, activities and crucial decision making abilities. However, due to the state of society as it stands in this country, mental health is often ignored and neglected. The current environment present in the universities of Bangladesh strongly indicates to a huge number of mental health cases that are naturally avoided and not taken into serious consideration. That is why, we believe that it is important to study the mental health of Bangladeshi university students and analyze exactly which factors influence the disruption of mental health. It is also important to pay attention to the type of university while researching, because many research factors are largely different in terms of public and private universities. In this research, we attempt to correlate a result of factors and association values that well suit our thesis purpose and also rightfully reflects on the students' mental states. It is our belief that if we can correctly identify and address the influential factors behind mental health turbulence, it will become easier to find out the mental health condition and how to combat those problems. This can also pave way for a collective support system for mental health patients in university students.

7.2 Future Work

From our research analysis, we have discovered, to a great extent, the prevalent factors that influence the mental health of university students and cause mental turbulence, leading towards depression. Our goal is to extend our research and work on ever greater data-samples, collected from all corners of the country and try to produce as accurate results as possible. We also hope to achieve greater results by incorporating more factors to study and test on, and using more in-depth models to find more correlation between those factors. We also wish to explore other mental health measurement scales such as USDI for more university-focused depression analysis. Furthermore, we can hope to apply our proposed research model, with necessary tuning, to measure depressive mental states in foreign universities as well in the future.

Bibliography

[1] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: Validity of a brief depression severity measure.," *Journal of general internal medicine*, vol. 16(9), pp. 606–613, Jul. 2001.

[2] M. A. Jamil, M. Sarker, and M. Abdullah, "Students' choice criteria to select a private university for their higher education in bangladesh," *European Journal of Business and Management*, vol. 4, pp. 177–185, Nov. 2012.

[3] K. Peltzer, S. Pengpid, S. Olowu, and M. Olasupo, "Depression and associated factors among university students in western nigeria," *Journal of Psychology in Africa*, vol. 23, no. 3, pp. 459–465, 2013. DOI: 10.1080/14330237.2013.10820652. [Online]. Available: https://doi.org/10.1080/14330237.2013.10820652.

[4] D. Sarokhani, A. Delpisheh, Y. Veisani, M. T. Sarokhani, R. E. Manesh, and K. Sayehmiri, "Prevalence of depression among university students: A systematic review and meta-analysis study," *Depression Research and Treatment*, vol. 2013, no. 373857, pp. 1–7, 2013. DOI: 10.1155/2013/373857. [Online]. Available: https://doi.org/10.1155/2013/373857.

[5] L. Ghaedi and A. MohdKosnin, "Prevalence of Depression among Undergraduate Students: Gender and Age Differences," en, *International Journal of Psychological Research*, vol. 7, pp. 38–50, Jul. 2014, ISSN: 2011-2084. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S2011-20842014000200005&nrm=iso.

[6] *Apriori algorithm in data mining: Implementation with examples*, https://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html/2, May 2015.

[7] D. Faeq, "Depression among students: Critical review," in. Dec. 2016. DOI: 10.13140/RG.2.2.21978.75205.

[8] P. A. Muller, "The effects of adversity on depression among university students," *Journal of Educational and Social Research*, vol. 6, no. 2, 2016, ISSN: 2240-0524. [Online]. Available: https://www.mcser.org/journal/index.php/jesr/article/view/9149.

[9] J. Brownlee, "Why one-hot encode data in machine learning?," Jul. 2017.

[10] O. Harrison, *Machine learning basics with the k-nearest neighbors algorithm*, https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d0176, 2018.

[11] A. Holstein, *I've experienced depression before — here are 5 things people get wrong about it*, 2018.

[12] "Suicide tendency rises among du students," *https://en.prothomalo.com/*, Nov. 2018. [Online]. Available: https://en.prothomalo.com/bangladesh/Suicide-tendency-rises-among-DU-students.

[13] M. M. Alsubaie, H. J. Stain, L. A. D. Webster, and R. Wadman, "The role of sources of social support on depression and quality of life for university students," *International Journal of Adolescence and Youth*, vol. 24, no. 4, pp. 484–496, 2019. DOI: 10.1080/02673843.2019.1568887. [Online]. Available: https://doi.org/10.1080/02673843.2019.1568887.

[14] A. Navlani, *Understanding logistic regression in python*, https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python, 2019.

[15] J. Brownlee, "Smote for imbalanced classification with python," Jan. 2020.

[16] J. Cerimele and L. Chwastiak, "Depressive disorder," in. Mar. 2020, pp. 6493–6498, ISBN: 9780198746690. DOI: 10.1093/med/9780198746690.003.0632.

[17] *Apriori algorithm in data mining: Implementation with examples*, https://www.softwaretestinghelp.com/apriori-algorithm/, Apr. 2021.

[18] M. A. Mamun, N. Sakib, D. Gozal, A. I. Bhuiyan, S. Hossain, M. Bodrud-Doza, F. Al Mamun, I. Hosen, M. B. Safiq, A. H. Abdullah, M. A. Sarker, I. Rayhan, M. T. Sikder, M. Muhit, C.-Y. Lin, M. D. Griffiths, and A. H. Pakpour, "The covid-19 pandemic and serious psychological consequences in bangladesh: A population-based nationwide study," *Journal of Affective Disorders*, vol. 279, pp. 462–472, 2021, ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2020.10.036. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032720328809.

[19] "Student commits suicide over failure to pass medical college admission test in dhaka," *https://www.dhakatribune.com/*, Apr. 2021. [Online]. Available: https://www.dhakatribune.com/bangladesh/2021/04/07/student-commits-suicide-over-failure-to-pass-medical-college-admission-test-in-dhaka.

[20] https://en.wikipedia.org/wiki/List_of_universities_in_Bangladesh.

[21] *A complete guide to the random forest algorithm*, https://builtin.com/data-science/random-forest-algorithm.

[22] *Bangladesh mental health*, http://www.searo.who.int/bangladesh/mental-health/en/.

[23] *Cambridge advanced learner's dictionary & thesaurus*, https://dictionary.cambridge.org/dictionary/english/university.

[24] *Decision tree learning*, https://en.wikipedia.org/wiki/Decision_tree_learning.

[25] *Depression*, https://www.who.int/news-room/fact-sheets/detail/depression.

[26] *Depression (major depressive disorder)*, https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007.

[27] *Depression in young people*, https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/Depression-in-young-people.

[28] *Everything you want to know about depression*, https://www.healthline.com/health/depression.

[29] M. Nawreen, *Why does ragging exist in our universities? answers and confessions*, http://api.hifipublic.com/2019/10/22/why-does-ragging-exist-in-our-universities-answers-and-confessions/.

[30] *Patient health questionnaire (phq-9)*, https://patient.info/doctor/patient-health-questionnaire-phq-9#.

[31] *Public vs private university*, https://www.studyinternational.com/news/private-university-vs-public-university/.

[32] *Random forest*, https://www.sciencedirect.com/topics/engineering/random-forest.

[33] *What is depression?* https://www.psychiatry.org/patients-families/depression/what-is-depression.

[34] Y. S. Won, D. Jap, and S. Bhasin, "Push for more: On comparison of data augmentation and smote with optimised deep learning architecture for side-channel,"

[35] D. S. M. Y. Arafat, "Enam medical college hospital," *9/3, Parbati Nagar, Savar, Dhaka, Bangladesh*, 774377982, 01716358146.