BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Inspiring Excellence

Predicting Depression in Bangladeshi Undergraduates Using Machine Learning

Authors
Ahnaf Atef Choudhury
Md Rezwan Hassan Khan
Nabuat Zaman Nahim
Sadid Rafsun Tulon

Supervisor
Amitabha Chakrabarty
Associate Professor
Co-Supervisor
Samiul Islam
Lecturer

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

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

December 2018

We would like to dedicate this thesis to our loving parents, families, friends
and Jalal Uddin Rumi ...

Declaration

It is hereby declared that this thesis /project report or any part of it has not been submitted elsewhere for the award of any Degree or Diploma.

Authors:

Ahnaf Atef Choudhury
Student ID: 15101022

Md Rezwan Hassan Khan
Student ID: 15101078

Nabuat Zaman Nahim
Student ID: 15301084

Sadid Rafsun Tulon
Student ID: 15101005

Supervisor:

Amitabha Chakrabarty, PhD
Associate Professor, Department of Computer Science and Engineering
BRAC University

January 2019

The thesis titled "Predicting Depression in Bangladeshi Undergraduates Using Machine Learning"
Submitted by:
Ahnaf Atef Choudhury Student ID: 15101022
Md Rezwan Hassan Khan Student ID: 15101078
Nabuat Zaman Nahim Student ID: 15301084
Sadid Rafsun Tulon Student ID: 15101005
of Academic Year 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of BSc in Computer Science & Engineering.

1. Md. Abdul Mottalib, PhD
Professor & Chairperson
CSE Department
BRAC University

2. Amitabha Chakrabarty, PhD
Associate Professor
Supervisor
BRAC University

3. Samiul Islam
Lecturer
Co-Supervisor
BRAC University

Acknowledgements

All praises are due to Almighty Allah, Most Gracious, Most Merciful who blessed us to be here at BRAC University for pursuing the bachelors degree.

Our journey towards the bachelors degree and in particular this thesis would not have been possible without the help of many people. It is our great pleasure to take this opportunity to thank them for the support and advice that we have received.

Our deepest gratitude goes to our Supervisor, Dr. Amitabha Chakrabarty, PhD and our Co-supervisor, Mr. Samiul Islam for their significant help and support throughout the course of this work. Without their continuous encouragement and guidance this thesis would not have been completed. We are truly grateful to have worked with such inspiring and friendly supervisors. We would like to thank Ms. Anne Anthonia Baroi and Mr. Shami Suhrid, Psychosocial Counselor and Lecturer, BRAC University for their valuable comments and feedback while designing the survey form and throughout the entire data collection and manipulation process.

We would also like to thank each and every one who were involved in the data collection process from our friends, faculty members, fellow classmates, juniors and all the respondents who took time out of their busy schedule to complete our survey. This research would not have been possible without their help.

We would also like to thank our family members, our parents, brothers and sisters who are the very reason of our existence. Without their unconditional love and support this research too could not be completed.

This thesis is dedicated to all of them.

Abstract

Depression is a major disorder and a growing problem that impacts a person's way of living, disrupting natural functioning and impeding thought processes while they might remain oblivious to the fact that they are suffering from such a disease. Depression is especially prevalent in the younger population of underdeveloped and developing countries. Youth in countries such as Bangladesh face difficulties with studies, jobs, relationships, drugs, family problems which are all major or minor contributors in a pathway to depression. Furthermore, people in Bangladesh are not comfortable in speaking about this illness and often misinterpret this disorder as madness. This research besides predicting depression in university undergraduates for the purpose of recommendation to a psychiatric, focuses on gaining valuable insights as to why university students of Bangladesh, undergraduates in particular suffer from depression. The data for this research was collected by a survey designed after consultation with psychologists, counsellors and professors. The survey was carried out through paper and Google survey form. The data was analyzed to find out relevant features related to depression using Random Forest Algorithm and then predict depression based on those features. A best method for predicting depression among Bangladesh undergraduates was found out after using six algorithms to train and test the dataset. Deep Learning was found to be the best algorithm with the lowest number of false negatives, closely followed by Gradient Boost Algorithm both with an F-Measure of 63%. Generalized Linear Model, Random Forest, K-Nearest Neighbor and Support Vector Machine were the other four algorithms used for comparison. The objective of this research is to determine reasons for depression and to check whether depression can be successfully predicted with the help of related features. Depression is an illness that people in Bangladesh tend to ignore and hence it builds up and worsens with time. This research aims to identify depression in its early stages and ensure a fast recovery for victims so that heartbreaking incidents like suicide can be avoided.

Table of contents

List of figures

List of tables

Chapter 1

Overview

A brief overview of the research addressing the problem and the aim, objective and method of solving it through this study has been presented in this chapter.

1.1 Introduction

The world is progressing at a rapid rate with the help of technology and human skills. Everyone is becoming so busy and materialistic that sometimes we forget to give some time to think about our mental health. To keep up with the rapid pace in the world people are constantly pushing themselves taking a lot of pressure both physically and mentally which have a adverse effect on their health especially their mental health. Depression is a silent killer which can harm a human being in a great way if not treated at the right time. It is a common mental illness and everyone at some point of their life is depressed. However, due to the lack of self-unawareness, the society and people's judgment this illness is considered as a taboo in many places across the world and people tend to make fun of this illness and tease people who are diagnosed with this problem. For instance, in Bangladesh, if someone has mental illness or is diagnosed with similar illness people assume that this person is mad. According to National Institute of Mental Health [59] the most common mental disorder is depression and around 16.2 million people experiencing depression. Depression is especially prevalent in students compared to the general population [64]. Research done in some low and middle-income country found out that students in universities, colleges, and medicals are more prone to depression and the result is an alarming signal that should be taken care quickly [45, 61, 35, 64]. Though this is one of biggest problem world is facing right now, people are not very vocal about it and little research been done to find out the root reasons of depression among students using data mining techniques, especially not in Bangladesh.

1.2 Problem Statement

Even though the world is moving forward and everyone is open to new things, mental illness is not welcomed in a good way in many societies. Especially in many Asian regions it resembles as taboo due to which people who are experiencing are not so vocal about their problem and try to hide their illness and try to overcome it by themselves. As a result, their illness may worsen which may deteriorate their health both mentally and physically and lead to severe depression which will directly trigger suicidal behavior and self-harm attitude in people. Along with this, most of the people are not aware about their depression issue and they don't take this so seriously like other illness believing that it will automatically heal if given time. On contrary to that some people become depressed thinking that they are depressed, but in reality, it may be not the case. While conducting this research we found out that most students act like that and were not so vocal about their problems, some of them think they are depressed which was affecting their performances but in reality, there were not depressed at all and some students don't even know that they have moderate and severe depression symptoms. Another problem we have found out that there is a group of students who are vocal about their depression problem and there is another group of students who are confused who want to know whether they are depressed or not so that they can seek professional help. However, the complex questionnaire's' in scaling methods and feelings of discomfort to give all the personal information to counselors demotivated them to do something about their mental illness.

Another huge problem we have found out during this research is that only identifying whether a student is depressed or not, will not help them to recover and heal from this mental illness. Finding out the root problems or reasons which are responsible for depressions in university going students should be one of the main priority. After a long interview session with BRAC University counseling unit and going through a lot of research we have discovered that even for counselor at first meeting it is difficult for them to identify the root problems which are causing depression in students. According to Ms Annie Anthonia Baroi, a counselor from BRAC University states that, for counselor, it is really difficult to know the real reasons to find out the real reasons why a student is depressed or not. This may happen because students sometimes are not comfortable sharing their all information with them and in most of the case the reasons which they did not share are the root reasons which cause depression.

So, there is no efficient platform so that students can determine whether they are depressed or not only just filling some basic questionnaires without feeling uncomfortable .Beside this

there is no particular way or research has been done to find out the root reasons which is causing depression in university going, undergraduate students.

1.3 Aim of Study

The main aim of this research is to find out whether a student is depressed or not using machine learning and data analyzing approach by only filling up some basic questionnaires which is related to depression instead of using scaling methods like The Beck Depression Inventory (BDI) to measure whether a student is depressed or not. Along with this we want to make it easier for mental health counsellor to find the root reasons and irrelevant reasons behind depression in university students so that they can understand students psychology more efficiently and give them the best advice and cure to their problems that they are facing. We have used 6 algorithms to run our model to predict whether a student is depressed or not. The algorithms are deep learning, generalized linear model, gradient boosted algorithm, random forest and support vector machine, k-nearest neighbor. We have used RFE known as Recursive Feature Elimination with Cross Validation and Random Forest Classification for selecting optimal features or reasons that is related to depression and eliminating irrelevant reasons that is not related to depression in university going undergraduate students which has been discussed in chapter 3 and chapter 4 of our paper.

1.4 Research Methodology

This research has been conducted in time span of 11 months and 19 days. However we are continuing with this research as we are planning to extend our research to discover more findings. We have started our work from 1st of December in 2017 which last till 20th November 2018. In first quarter of December we spent time on brainstorming about our ideas and we took around 1 month to come up with our initial idea.After that we started to read similar articles that is related to our idea. At that phase of time our idea was still in development process.At the same time we met with many counsellor to know more about about depression and its related work. They were so helpful to us and helped us throughout whole time till the last stage of our thesis.Taking guidance from our supervisor and BRAC university counsellor unit we were able to finalize our idea and we started our studies on depression,scaling method to find out depression and did research to find our relevant reasons and features which are related to depression in undergraduate university going students which start from January 7th 2018 which last till May 1st 2018.We were really careful when we were preparing our survey form and we took took around 3 months to make our survey form.From May 20th to May

30th 2018,we learned the formal process of how to take a survey that is related to mental health from our respective faculties from physiological department of BRAC University.

Taking all the precaution and with formal planning,we started to take our survey from BRAC University from 6th of July. Along with that we also started to take survey from other universities which we conduct through online surveys. Before taking online survey,we made a video which clearly states what our thesis is all about and clearly mention all the rules and procedure to fill up the survey form. Overall we take around 3 months to complete our first phase of data collection process which last till 30th September though we are still collecting data from online survey.

After we are done with data collection,it was time for data cleaning and data pre-processing before we applied machine learning mechanism on it. So from 1st August we started our data cleaning process which last till 1st of September. In total we had 7 versions of data sets during our data cleaning process from which we came up with our final data set that will be used in machine learning mechanism. Afterward for whole 1 month we applied our machine learning algorithm to our processed data set considering all the necessary parameters which was needed in machine learning mechanism.In parallel we started writing chapter 1 and 2 of our thesis paper from 1st of September. Afterwards, we wrote chapter 4 and 5 which took around 1 month to complete. And at last we finish the remaining part of our writing the thesis paper before November 8th 2018. We took around 14 days to check our paper thoroughly and then convert out writing into the given format of BRAC university with the help of latex software. The entire details is shown in Figure 1.1

1.5 Thesis Outline

- In chapter 2, brief introduction about prevalence of depression among university going students and how to find depression through popular scaling method ,background studies on machine learning algorithms and previous work related to this research has been discussed
- In chapter 3,detailed description of data collection process,discussion on relevant reasons of depression in undergraduate students ,explanation of data cleaning and reprocessing of sample data and lastly illustration of data visualization of the final data
- Chapter 4 represents detailed explanation of the experimental result and analysis which includes process of applying algorithm,cross validating the data sample ,finding out

the accuracy for different algorithms and detail description of process of finding out the optimal features that can lead to description.

- Chapter 5 concludes the paper with general remarks,limitations,improvements and future works of our research.

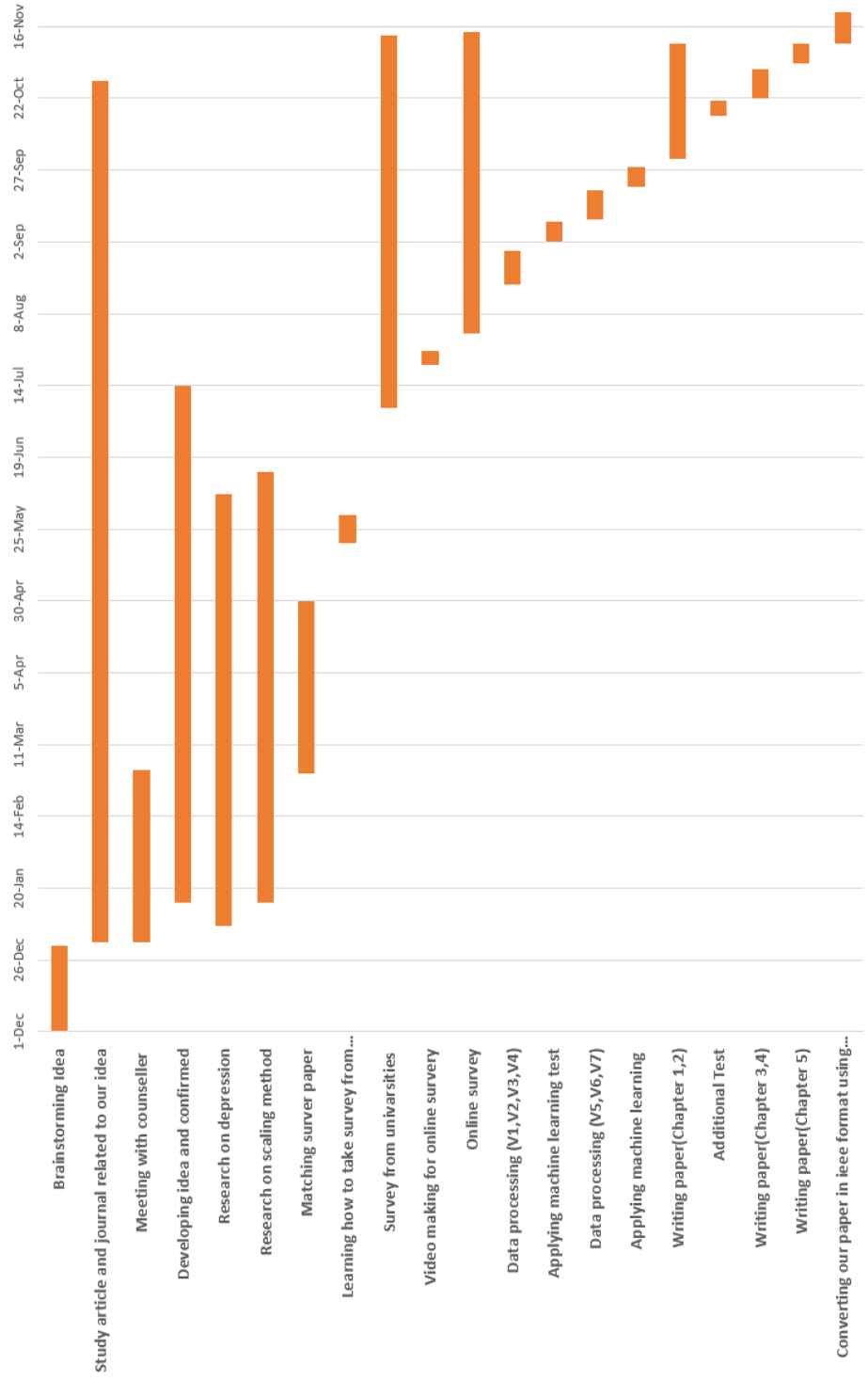


Fig. 1.1 Time span of research

Chapter 2

Background Study

This chapter further describes the problem with respect to our sample population, the scaling methods used and presents reviews of similar work in the literature.

2.1 Discussion about depression

Depression is a common term that we use often when we talk to family and friends and say that we are depressed. However, not all feelings of sadness and grief can be classified as depression. The American Psychiatric Association describes depression (major depressive disorder) as a common and serious medical illness that negatively affects how we feel, the way we think and how we act [55]. According to them, in any given year about one in fifteen adults (6.7%) can be estimated to be affected by depression. Also, 16% or one in six people encounter depression at some point in their life. Studies also show that people usually experience depression around the age of 20 years which is one of the prime reasons why we chose to conduct our research among undergraduate university going students. Depression is more prevalent in women with studies suggesting that about one-third women suffer from an episode of depression at least once in their lives.

Sadness and grief may arise several times in life due to loss of a loved one, problems in professional and personal life or due to no specific reason. However, depression usually makes people feel worthless consequently making them hate themselves. Our research focuses not on general cases of people feeling down and unhappy nor on medical conditions such as thyroid problems, brain tumor or vitamin deficiency that causes symptoms similar to depression. We instead look at actual cases of people with psychological and clinical depression who have been diagnosed or need diagnosis. The various symptoms of depression range from persistent feelings of sadness and anxiousness, pessimism, loss of

interest, laziness to difficulties in sleeping, weight changes and in extreme cases thoughts of suicide. We have considered all of these reasons in preparing our survey, some in the general set of questions as features for machine learning classification and others in the two depression scales. Details about these have been presented later in the paper. For a person to be diagnosed with depression, these symptoms must be present for at least two weeks [3]. Health line, an US-based leading provider of health information says that only profound sadness that persists for more than two weeks and restricts us from functioning properly may be a sign of depression [57].

According to psychiatry department at Harvard Medical School, the four most common types of depression are major depression, persistent depressive disorder (formerly known as dysthymia), bipolar disorder, and seasonal effective disorder [48]. A brief description of each of these is given [3, 57, 48, 33, 67]. Major depression, more widely known as clinical depression is the most common type of depression where extreme sadness lasts for a long time (weeks or months) with a complete loss of interest in even pleasurable activities among other difficulties. Persistent depressive disorder or dysthymia is an episode of depression less severe than major depression but lasting for two years or more where depression levels may become less or more intense at times. Bipolar disorder consists of periods when a person is very happy and proactive (hypomania) followed by a period of depression. Angelos Halaris, MD, PhD, a professor of psychiatry and the medical director of adult psychiatry at the Loyola University Medical Center in Chicago says that although the bipolar disorder affects only 2 to 3 percent of the population, it has one of the highest risks for suicide [33]. Seasonal effective disorder (SAD) occurs due to seasonal changes, mostly due to cold and dark winter days. In countries like Bangladesh where winter is short and irregular, seasonal depression is not so common although no studies have been conducted in this regard. Apart from these depression types unique to women include perinatal depression (occurs during pregnancy or after delivery) and PMMD (depression due to premenstrual syndrome). Psychotic depression resulting from mental health problems is not generally associated with depression. Although we did much background study on the types of depression, identifying the types of depression associated with the people in our data set was not in the scope of our study. We however hope to conduct future research in this regard with a larger data set.

2.2 Discussion about depression in university students

According to the Center for Collegiate, Mental Health report depression and anxiety are the top reasons student are seeking counselling for and it shows that 1 out of 5 university

students are depressed. University students are more depressed than the general population [32]. According to another research, the depression rate in under-grad university student is high in the developed and underdeveloped country where income is basically low [35]. A number of students in college raking anxiety and depression counselling are increasing day by day [61]. Most of them either attempted suicide or try to harm themselves, and some of them find it was difficult to work properly and some of them discovered severe anxiety issues.

According to the article [45] "Prevalence of depression and its associated factors using Beck Depression Inventory among students of a medical college in Karnataka" their result suggests that Depression is highly prevalent among medical students. According to the research , it states that 71.25% people are found to be depressed among 400 students and among this 80% had mild and moderate depression ,7.5% had severe and 6.7% had profound depression.

2.3 Scaling method to determine depression

A depression rating scale is a standard measurement instrument to analyze a person's behavior in order to determine whether further checking is required for that individual to be diagnosed with a depressive disorder. The depression scales consisting of various questions and inquiries to be filled up by subjects under study are often so accurate that they can identify or predict depression in people although the purpose is more to recognize people at risk of developing the disorder. Depending on the severity of depression as determined by a depression scale, psychologists or experts decide whether or not an individual should be subject to further observation to find out if he/she should be diagnosed with depression. There are several rating scales that serve this purpose. We went through extensive study and research including reading articles, conference, papers, journals and meetings with professors and psychologists from BRAC University and Dhaka University before choosing the Beck Depression Scale and (Bangla Depression Scale) as the instruments for measuring depression levels of the students who took our survey. The process in which we came up with these choices is briefly described below.

Depression scales can generally be classified into scales completed by researchers and scales completed by patients. Two of the scales completed by researchers, the Hamilton Depression Rating Scale and the Montgomery-Åsberg Depression Rating Scale are typically used for assessing the effects of drug therapy hence were eliminated instantly. The

Raskin Depression Rating Scale rates verbal reports, behavior, and secondary symptoms of depression. All of these scales were not considered principally because of time constraints as individual patient interview and monitoring of all subjects was not feasible in the scope of this work.

Among the most popular of the scales completed by patients is the Beck Depression Inventory (BDI, BDI-1A, BDI-II). The Beck Depression Inventory (BDI) is one of the most widely used screening instruments for measuring the severity of depression in both adults and adolescents over the age of 13 (McDowell Newell, 1996) [82]. The Beck Depression Inventory is the most widely used instrument for detecting depression [2]. Before the advent of the Beck Depression Inventory, rather than considering that this disorder could be resulting from a patient's thoughts, mental health professionals looked at depression from a psycho-dynamic perspective. The most recent version of this scale BDI-II, published in 1996, is designed for individuals aged 13 and over with it components covering all the major reasons for depression especially in young people. The Beck Depression Inventory is a self scoring scale that consists of 21 multiple choice questions relating to various symptoms and causes of depression. Each question has four choices in increasing intensity of scores from 0 to 3. After completing the questionnaire, the scores of all the questions are added to measure the severity of depression if depression is at all present. There are six categories of result with total points being below 10 considered as normal and aggregate score of over 40 being classified as extreme depression. A sample of the Beck Depression Inventory is given at the end for reference.

Other scales completed by patients include Geriatric Depression Scale (GDS), another self-administered scale which was ruled out as a choice for use in our research since it was more appropriately used to analyze depression in older populations particularly patients with traces of dementia. The Zung Self-Rating Depression Scale is another simple scale similar to the Geriatric Depression Scale in that the answers of all the 20 items were either "a little of the time", "some of the time", "good part of the time" or "most of the time". The GDS has answers that are simple "yes" or "no" suitable for older patients. The Patient Health Questionnaire (PHQ) although more relevant, is very short with the Patient Health Questionnaire-9 (PHQ-9) composed of only 9 questions. Furthermore the scale specializes in measuring all form of mental disorders and is not limited to depression only which is a major limitation from our perspective. The Hospital Anxiety and Depression Scale (HADS) besides being less used is too short and simple to consider the complicacy of young adults. We considered using the Center for Epidemiologic Studies Depression Scale (CES-D) for a

while but eventually found it to be less reliable and less sophisticated for use in a university undergraduate student population. Most of these scales for measuring depression were not chosen either because they were irrelevant to the age and behavior of our dataset population or because they were too simple to be used in practice for research purposes especially not for predicting depression in individuals based on a few simple features.

The depression scale in Bangla was prepared by SM Abu Hena Mostafa Alim for their own research titled "Translation of DASS 21 into Bangla and Validation Among Medical Students". This scale was selected and used particularly because of its relevance to the climate and culture of Bangladesh. In spite of the fact that a few questions on both the scales were similar, the Bengali scale contains certain lifestyle related questions associated to our way of life that a depression scale designed for the world fails to consider. These small subtle differences might reveal something about depression unique to the people of Bangladesh. Results using this scale could also be used to compare relevancy of the internationally used scale from the perspective of Bangladesh and also to capture redundancies in the information supplied by the person taking the survey in terms of whether misinformation on the part of the respondent gives different results in the two different depression measurement scales. Setting aside the benefits of comparative study, the Bengali scale was principally included as a respect to the language for which 3,000,000 people gave their lives and between 200,000 and 400,000 women were raped in the Bangladesh Liberation War of 1971.

2.4 Related work to determine depression using scaling methods

In the paper [17], the researchers made use of data mining, more specifically classification to predict possible depression in people in the future if they are not already suffering from it. They used synthetic data prepared with the help of a Java program to carry out the research. The training sample consisted of 600 instances whereas the test set had 400 instances of data. The attribute selection procedure carried out by the researchers was an intense and thorough work in which many online surveys and questionnaires were analyzed. They selected 31 attributes, the last one was the class variable “May Have Depression”. The attributes could take values from 0 to 3 based on the severity of the symptom with 0 being ‘None’ meaning the attribute was absent in that person and 3 being ‘Serious’ meaning the person was severely suffering that symptom with ‘Mild’ (1) and ‘Medium’ (2) in between. We carried out a similar process for our research on analyzing depression among university going students

where we did extensive research on why teenagers and young people suffer from depression and which method or depression scale would be best in measuring depression for them.

The study used the popular machine learning and data mining tool WEKA for classification in order to find out hidden patterns in the data. They used the C4.5 decision tree algorithm to get the classification model. An improved version, J4.8 was used to find out the results of the depression classification model. In the training process, out of the 600 instances, 555 were correctly classified giving an accuracy of 92.5 percent. Their confusion matrix showed 263 true positives, 292 true negatives, 11 false positives and 34 false negatives. Accuracy, precision and recall were the classification metrics used by them for evaluation. They then validated their model using the training data. In this phase, 333 instances out of 400 were correctly classified giving an accuracy of 83.25 percent. The confusion matrix with the training sample had 163 true positives, 170 true negatives, 27 false positives and 40 false negatives.

They later used the depression model with 20 new data instances to find out the diagnosis of these unknown cases of depression. The model classified 13 as having no depression and 7 as 'yes' meaning they were future possible depression cases. The researchers also used WEKA to represent the results in a different way using probability distribution which showed that ten of the twenty classes were predicted with a probability of 1 meaning that the prediction was absolutely correct. Moreover, the smallest probabilities of predicting a 'yes' and 'no' class was 0.829 and 0.778 respectively clearly illustrating that the predictions were accurate and could be trusted. The study mentions that the probability distributions could be helpful to physicians and could be used in other medical applications of data mining.

The article concludes by discussing the importance of selecting a large attribute set as they did particularly in depression related studies since depression and other somatic illnesses have similar symptoms and it is quite difficult to distinguish between them. We considered this factor in our own research which is why we chose the "Beck Depression Scale" which took into account many of the symptoms of depression in their long but simple 21 questions. Our own study diverted from the approach of using synthetic data to test and validate model and focused on collecting real data by preparing a survey after much thought, consultation and research and using that data to train and test the model which is exactly what we have done. We have the opportunity to try out the performance of our model using synthetic data too, later on.

In another paper [8], researchers worked out a best method for predicting depression among

older people using machine learning classifiers. WEKA, a data mining tool developed by the University of Waikato in New Zealand that can apply different machine learning techniques on problems like data pre-processing, Forecasting, Classification, Prediction and Regression was used for the research. By comparing results, a best method to predict depression was chosen.

The study proposed an automated system to tackle depression which is quite prevalent among senior citizens, by considering various socio-demographic factors and co morbid condition. Their aim was to replace problems associated with manual diagnosis and treat patients as early as possible. The dataset used for training was collected from a slum at Bagbazar, Kolkata, a service area of Bagbazar Urban Health and Training Centre (UHTC). Sixty senior citizens aged minimum 60 years were interviewed using Geriatric Depression Scale (GDS). Five classifiers, namely BayesNet Classifier (BN), Logistic, Multilayer Perceptron (MLP), Sequential Minimal Optimization (SMO) and Decision Table (DT) were compared with respect to four metrics which are Accuracy, ROC area, Precision and Root Mean Square Error (RMSE). The test data with no decision class was created with data of ten more citizens. They then checked the predicted result manually with the GDS scale.

Supervised learning was used in the learning process to predict output of test data was used since output label of training data was known. After the input data set was loaded in WEKA, they followed these steps. Supervised filter – attribute Selection was chosen for pre-processing following which the five classifiers are chosen and experiment was run 5 times with classification output being recorded each time. The researchers utilized all three test options available in WEKA. The simplest of them, “Using Training and Testing Set” just trains network using training set and then tests the network on a test set. The second option, “Cross-validation” is quite interesting and critical where the test set is generated automatically from training set and number of folds is provided to prevent over fitting. For example, if number of folds provided is 20, data is divided into a training set containing 80 samples and a test set of 20 samples. The next epoch is done using the test data that is prepared from the previous 80 samples and the 20 samples which were previously used as test data were now part of the training data. The final result is obtained by averaging the results of the different epochs. In addition they also used a third test option percentage split where the data set is split into training and test data according to percentages specified by the user.

They found out using the training and test set option that SMO with an accuracy of 93.33

percent and a precision of 0.94 was the best prediction model closely followed by BN with an accuracy of 91.67 and a precision of 0.92. However, when the other two metrics ROC Area and RMSE was observed BayesNet with an ROC Area of 0.98 and RMSE of 0.25 seemed better. Again, when 10 folds cross validation was used, the researchers recorded a highest accuracy of 88.33 percent and a maximum precision of 0.88 from SMO but the result was once again contradicted with BN having a much higher ROC Area (0.96) and a lower RMSE (0.32). The consistency with respect to all metrics was observed using percentage split option where BayesNet with accuracy, precision, ROC Area and RMSE of 95 percent, 0.95, 0.99 and 0.22 respectively was better than the other four classifiers and hence chosen as the best classifier with the percentage split appearing to be the best test option according to this research.

However, factors such as association of other problems to depression which may mean that someone is not actually depressed could affect results. Furthermore, different nature inspired algorithm based optimization techniques can be used for more accurate feature selection of depression prediction. Also, the dataset can be made using data from different parts of the country and if possible other countries as well to observe consistency of results. In our own research, we used a tool quite similar to WEKA, named RapidMiner [28]. RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization. The researchers in the study also mentioned the lack of work on depression prediction among senior citizens in the literature which motivated us to do something similar. Hence, after much thought and research we decided to work on analyzing depression patterns among the younger generation, in particular, university going students. There was even lesser work on depression among teenagers than was among senior citizens and close to no work on the depression scenario of our country, in particular, especially not university going students. Hence we wanted to combine data mining and machine learning to predict as well as find out the reasons for depression among university going students in Bangladesh.

Researchers in another study [50] report how people suffering from depression think deeply and repetitively and express this same behavior on both Twitter and their real lives. In particular, thoughts about sleep, pain and suicide were taken into account. They carried out

the research by considering a group of people who had been diagnosed with depression as signified by their tweets over the course of a week. Their Twitter data was compared with the Twitter data of a randomly chosen control group who as per their Twitter history had never suffered or been diagnosed with depression.

The study was done in the following way. At first, some patterns and combinations of keywords reproduced from previous research on depression was used to identify 316 unique Twitter users with depression. Care was taken to ensure that tweets by support groups, commercials and re-tweets were not taken into consideration. Then the full Twitter histories of these 316 individuals were downloaded to look for symptomatic characteristics (sleep and pain) and dangerous thoughts (suicidal thoughts). "NVivo 10 and NCapture were utilized to extract user time lines via Twitter API. MS Access 2013 was used to store aggregated tweets and Hyperion Interactive Reporting Studio 11 was used to search and analyze the tweets for tweet contents." Similar approach was used to form a control group of another 316 Twitter users who had never been diagnosed with depression according to their Twitter time line. The data from the Twitter histories of all users was dissected to look for the two categories of rumination using words or combinations of words determined by research on depression.

Fisher exact test, an independent t-test, Chi-square test all showed a higher proportion of Tweets about sleep, pain, suicidal thoughts by the depressed/study group than the control group. This however is only true because they assumed that the no one in the control group is depressed just because they have not Tweeted about it. This assumed seems quite far-fetched particular given that the nationality, race, gender and some other factors were not analyzed for a link with depression in the study. The paper also does not reveal whether the Twitter users were all from a particular university, region, and country or not. The researchers however, highlight the fact that not considering the presence of other mental illness among individuals from both groups could indeed limit the scope of their findings hence requiring further research and analysis.

2.5 Related work to determine depression using machine learning

This paper[10] explained how they predict whether a twitter user is suicidal or not using machine learning on basis of validate self-report which is used to find suicide rate and with the tweet feeds/status of active twitter users. The authors main work is to predict whether a

user is suicidal or not using their tweeter feeds and this results are then cross checked with the result of self-report the users filled previously for this research purpose.

The data for this research is collected very carefully. A survey named "Survey for Twitter user" is taken where all the participant were active twitter users from USA. They were given 3 sets of questionnaires. The questionnaires are Depressive Symptom Inventory–Suicide Subscale (DSI-SS), The Interpersonal Needs Questionnaire (INQ), Acquired Capability for Suicide Scale (ACSS) to assess suicidal rate. In their questionnaires they included five control questions to make the data collection more reliable. These control questions helped to identify users who did not answer or fill the questionnaires carefully and they excluded 46 more participants who was not very careful when doing this survey and their final sample data size was 135 on which they run their research. The sample data contain 85 females and 50 males from 6 different ethnicity, education income, and twitter account creation date. After that article explain the process of analyzing the tweets of each user. Each user tweets are retrieved aggregated in a single file. Each file of individual user analysis by updated 2015 version of Linguistic Inquiry and Word Count Software (LIWC). It software which is can correlate with sucidality in the context of social media.

Afterward article described how to predict their expected result which is to predict whether a user is suicidal or not using their tweeter feeds. They use decision tree as their predictive model and they implemented the predictive analysis in Python, using the scikit-learn library. Author use leave-one-out-cross validation for estimating the accuracy of the decision tree learning. The loo-cv accuracy was 91.9%.[(9+115/135)*100)]. False positive is 0.75 and false negative is 0.93. It accurately identify 9 result as suicidal and they were really suicidal in real and accurately identified 115 individuals as non suicidal and they were non suicidal in real.

Firstly, twitter feed is not the best measure to asses suicidal behavior as the reasons are not specific and it may raise question on the validity and logical approach of this research paper[10]. Secondly instead of doing the research on the same age group of people they have done their research on various age group of people.

The paper[56] explained that using clinical variables related to suicide and using demographic variables it can predict whether a person will attempt suicide or not using a machine learning approach. They have used 3 algorithms and they are LASSO, SVM, RVM and it is implemented in MATLAB. The overall accuracy range was between 65%-72%. Among

them, RVM gave the best accuracy result of 72%.

This study was done on 144 subjects and they collected the data very carefully so that they can get accurate information from the user. A lot of information about each subject has been taken for this research purpose. The set of information are demographic histories, Axis-I diagnoses and clinical characteristics were assessed using the Structured Clinical Interview for DSM-IV axis-I Disorders (SCID-I. Current dimension mood and anxiety symptoms were asses using the Hamilton Depression Rating Scale (HDRS), the Young Mania Rating Scale (YMRS), and Hamilton Anxiety Rating Scale (HARS).

So in this way, they had identified 15 predictive variables upon which they trained their model to find whether a person will attempt suicide or not.. All these are categorical variables which are normalized by z-scoring where it labels 0 as no and 1 as yes except one variable which is a continuous variable.. They use cross-validation method LOOCV for evaluating their model accuracy. The validity of the algorithms in predicting individual suicide attempters from non-attempters was evaluated using prediction accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). the accuracy of 3 models ranges between 64.7 to 72%. Among these RVM gave the best result which is about 72% and SVM gave 64.7% and LASSO gave 68%. RVM was able correctly identified 103 out of 144 subjects as either suicide attempter or non-suicide attempter.

The main thing which we have taken from this two papers[10, 56] are the data collection method and how they have predict whether a person is suicidal or not using their tweeter feeds and machine learning approach and their machine learning approach and which algorithm they have used. As their paper is related to our work , their algorithm worked best in our data too. We have RVM, SVM and Lasso which we get to know from their paper. Firstly, we have made a screening system of our data collection after having reading about their idea. In section 2 we have 16 questions of our survey from. Secondly they collected data of Depressive Symptom Inventory–Suicide Subscale (DSI-SS), The Interpersonal Needs Questionnaire (INQ), Acquired Capability for Suicide Scale (ACSS) to assess suicidal rate result which the twitter user had to filled up to predict whether a person is suicidal or not[]. Then using their twitter feeds they trained their machine learning model to find that whether they are really depressed or not comparing it with the actual result. In our paper we have find out whether a student is depressed or not using the beck depression and Bangla depression scaling method . Then we tried to predict whether a person is depressed or not using relevant reasons that may lead depression. Thirdly they were able to predict the most important

reasons of suicides. So in our paper we have shown that what reasons are relevant and most important that lead to depression and what reasons are not important and not related to depression.

2.6 Machine Learning

2.6.1 What is Machine Learning?

Learning is basically converting experience into knowledge or expertise. In a machine learning algorithm training data represent experience and is given as input into the algorithm to train it and learn it and the output resemble knowledge or expertise which is learned with the help of training data [68]. Mathematically it is to search for the best possible sets of function h : X $\rightarrow$ Y [47]

Domain set: This is the set of object we want to label(X).Domain set can be represented by vector feature. For example to determine papaya is tasty or not vector features can be papaya color and softness, Label set: This is the result what an algorithm gave to a given problem . For example machine have to predict whether a papaya is tasty or not.So ,let Y be 0, 1, where 1 represents being tasty and 0 stands for being not-tasty. Label set can be non binary as well.

Training data:: S = ((x1, y1). . .(xm, ym)). This is given set of input and its corresponding output which is used to train the model. For example x1 is a papaya and y1 is the result of that papaya being tasty or not. So is x2,y till sm,y

The learner's output: h is a function known as the predictor which is the best possible set of function that can predict the result as close as accurately. h : X $\rightarrow$ Y ,searching out the best possible set of predictor where X is the set of possible inputs and outputs set of possible outputs Y, for those inputs.

Measures of success: It is the rate of error not predicting result correctly. For the papaya example it is the rate or error not predicting correctly whether a papaya is tasty or not [68].

2.6.2 Supervised Learning

Our research is based on supervised learning. This is because we have both inputs and outputs [47]. In this learning training data consists of pair of input and output.S = ((x1, y1). .

.(xm, ym)) where si is the training set containing sample input and output use to train the model. In supervised learning label is known. Identifying spam mails and separating it from the mail which is not spam. Training set of this will consist of spam and non spam mail and algorithm will learn which is a type of spam mail and which is not. In this way it can train a model using this training set to differentiate future mail which is spam mail and which is non spam mail [68].

2.6.3 Algorithm used

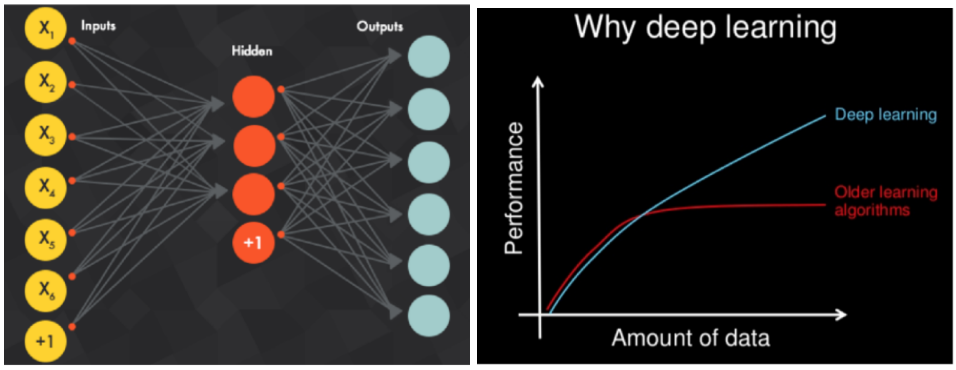
We have used 6 algorithms to determine depression. A brief idea about the algorithm has been given.

- **Deep learning**

  Deep learning is a sub-field of machine learning that is inspired from the structure and function of the brain called artificial neural networks [13, 25, 38, 69]. As shown in 2.1a, this algorithm enables machines to learn from previous experience using hidden layers in the neural network. There has to be at least 3 layers. A lot of data needs to fed to the computer system which can then be further used to make other decisions. It will gradually learn by using simpler graphs and then updating itself by going deep into the hidden layers of the artificial neural network. One big example of deep learning is the autonomous car. The autonomous car's navigate using the sensors and on board analytic which are then used to recognize obstacles and react to them appropriately. From figure 2.1b we can see, deep learning compared to other traditional machine learning techniques can handle lots of data thus providing better performance for more data. Whereas for other machine learning techniques, after a certain data is fed to the machine, performance becomes constant.

  Furthermore we can see from figure 2.2 that in deep learning the algorithm itself does feature extraction whereas for the other machine learning techniques the humans manually extract features.

  Activation function are used to decide the output of a neural node from given inputs into the nodes. Activation function is represented by f(h) where f(h) is 0 if h is less than 0 else it is 1. Here h is input of the output unit. Initially, weights are initialized as random which is multiplied with the input to a neuron. As more data is fed into the system and as the machine matures over time, it learns how to classify. As a result,



(a) (b)

Fig. 2.1 (a)Artificial Neural Network [38]; (b)Better performance [13]

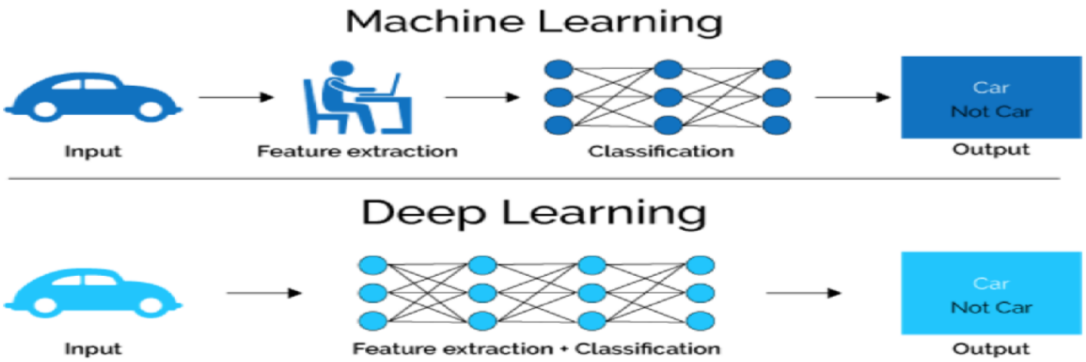


Fig. 2.2 Auto feature extraction [38]

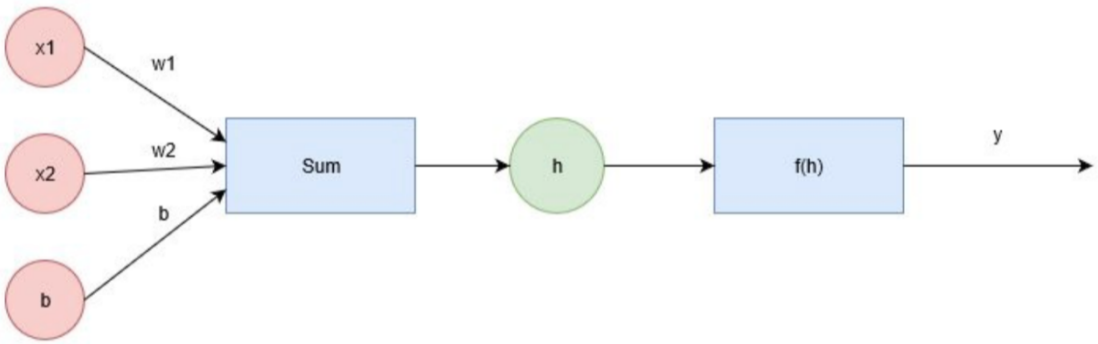


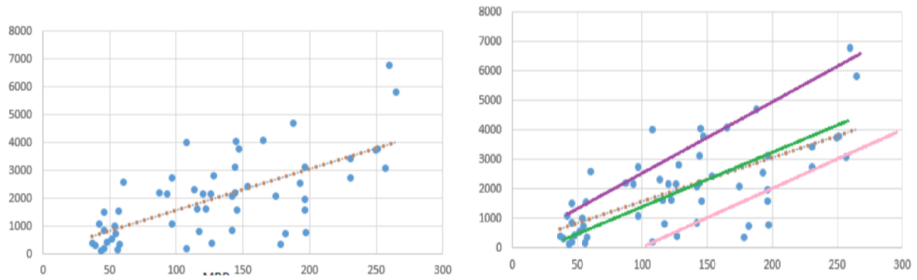
Fig. 2.3 Output y from weights, inputs and bias [38]

the weights are adjusted based on any errors in categorization learned throughout the training period. In the following Figure 2.3, we can see how the weights, inputs and

bias form the input h which is further passed through activation function to get the final output.

- **Generalized Linear Model**

Generalized Linear Regression is a machine learning algorithm for both classification and regression technique. It is one of the simplest algorithms since it uses statistical method for prediction and hence it is widely used by the data scientists all over the world [51, 34, 5]. Basically this algorithm works on the basis of independent and dependent variable. For a single instance of data, we have n number of features and all these features are independent variables and the dependent variable is the final outcome. We then assign weights to each of the features based on the importance of each of the features. For example, if a student wants to get a good grade, then he/she needs to be a regular student, which has more importance and hence will get the highest weight.



(a) (b)

Fig. 2.4 (a)Actual values of predicted lines [34]; (b)Selecting the best fit line [34]

Lets assume we have one feature for the model. Here number of classes attended is the independent variable and result is the dependent variable. Therefore our linear regression model will have a straight-line equation. From Figure 2.4a we can see how to plot a graph for all the instances of our training data and then draw a best-fit line on the graph for prediction. Now, since the data is scattered through the graph how do we select which is the best-fit line. There can be multiple predictive lines as we can see in Figure 2.4a.

Therefore to predict the best possible line we find the residuals. The main purpose is to have a line where our predicted values should be closer to the actual values. So, we need to reduce the distance between our predicted value and the actual value and this is

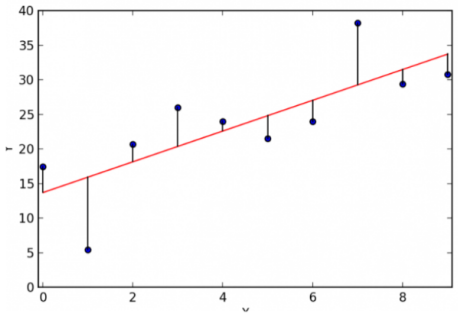


Fig. 2.5 Residuals of actual predicted values [34]

known as the error. This error is also termed as residuals. The vertical lines shown in the figure 2.5 represent the residuals. We then find the sum of the square of residuals.

- **Gradient Boosted Tree**

Gradient Boosted Trees is a powerful machine learning technique for classification and regression problems to make a very good predictive model in the form of an ensemble of weak prediction models [23, 12, 26]. This algorithm is used to change the weak assumption done initially to a very good assumption. From Figure 2.6, Figure 2.7 and Figure 2.8 we can see that the algorithm runs several times sequentially so that it can update itself from the previously found examples that were wrongly classified. Therefore, it can be said that boosting is a method in which predictions are not made independently but rather sequentially.

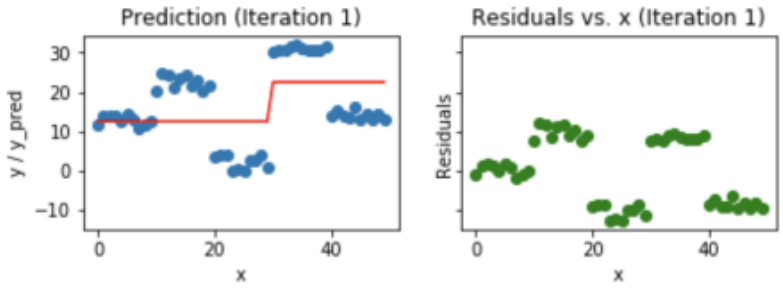


Fig. 2.6 First prediction done using the algorithm [26]

From Figure 2.7, the method learns from the mistakes of the previous predictors. Hence, the observations predicted by the previous predictors thus have unequal probability to appear in the next models and hence it is based on the error done by the models. For

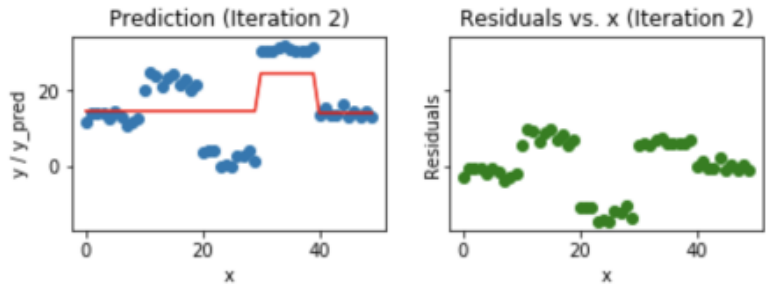


Fig. 2.7 Learning and predicting from previous mistakes [26]

this we need to find the Mean Squared Value (MSE), which is called the loss function. Less the loss function is, better the prediction model will be the. We get the minimum loss function using gradient boosting technique and updating through the repetitive predictions.

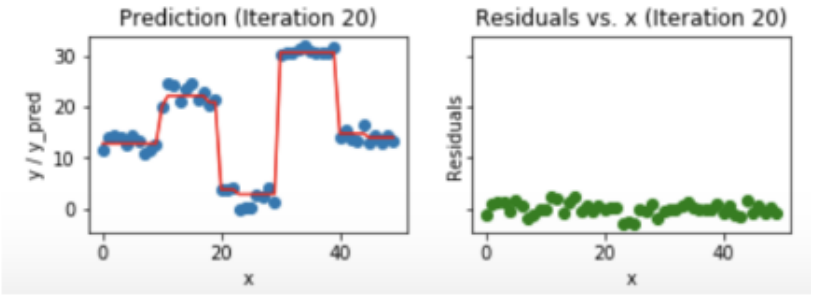


Fig. 2.8 Certain iteration where best prediction is made [26]

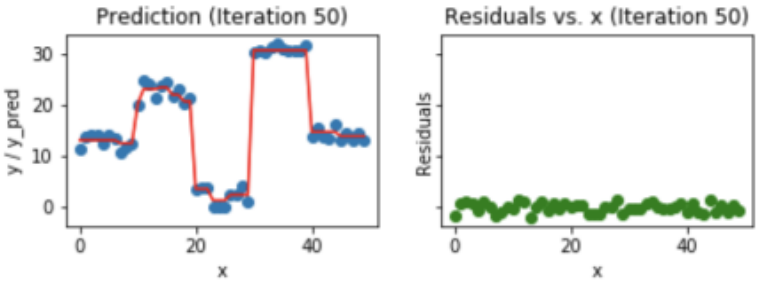


Fig. 2.9 Over fitting if stopping criteria not selected [26]

However, we have to be very careful at choosing the stopping criteria shown in Figure

2.8. By this we mean that we have to choose when to stop the algorithm so that we could avoid over fitting of the training data as we can see in Figure 2.9.

- **K Nearest Neighbor**

K Nearest Neighbor (K-NN) is a non-parametric, lazy learning algorithm that is used for both the classification and regression problems [39, 11, 72, 74]. By non-parametric we mean that it does not have any knowledge on the data distributions. By lazy we mean that it does not classify based on the training data points. As a result, it is not trained using any specific training model for classification. All these training data points are needed during the testing phase. This is because K Nearest Neighbor algorithm works on the basis of the nearest neighbors of the given data point and then predicts based on the majority of the neighbor's classification shown in Figure 2.10. For this reason, K-NN works best on smaller data sets with fewer features.

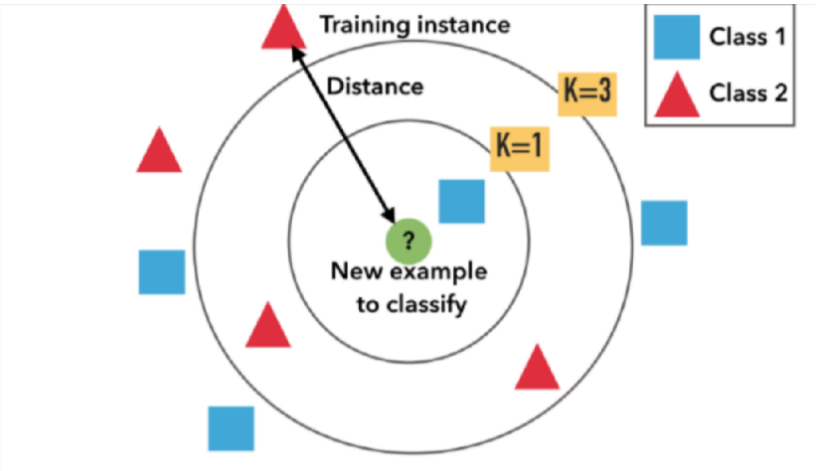
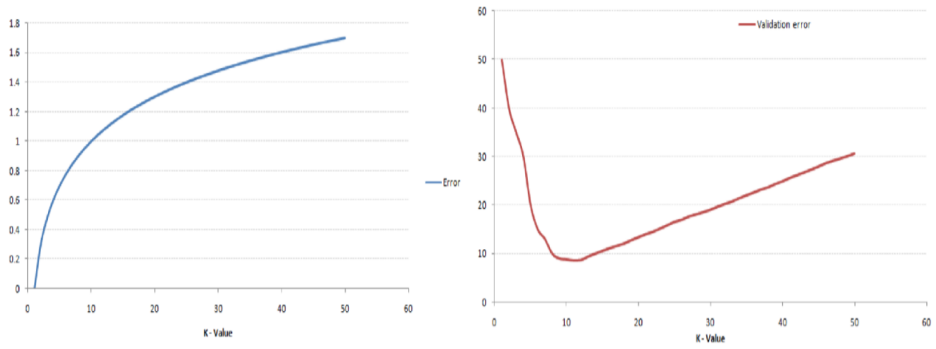


Fig. 2.10 Prediction based on nearest neighbor [11]

Now as we can see in figure 2.10, classification for the algorithm will vary depending on K. Here, K means the number of neighbors we take from the training data for classification. When we select the data point for classification, based on K a circle will be drawn with data point as the center of the circle. Inside that circle K number of nearest neighboring points will be present. Now, based on that maximum number of classifications of the testing points, final outcome will be predicted.

Now how do we choose the value of K. Well that depends on the training error rate and the validation error rate. For the example in Figure 2.11a, we can see that if k



(a) (b)

Fig. 2.11 (a)Training error rate [74]; (b)Validation error rate [74]

equal to 1 then we have 0 error rate and hence it seems that our value for K would be 1. However if we look at Figure 2.11b, the validation error at k equal to 1 is very high. This shows that we should not take the value of K as 1. Therefore to sum up we can say that based on the validation and error rate of the training data set we will choose the value of K.

- **Random Forest**

Random forest is one of the most widely used machine learning algorithm since it is very easy to use and implement. It is highly used for both classification and regression [46, 20, 44] . For our case we will be discussing about classification done by random forest. It is supervised learning technique where we feed labeled training data to the algorithm and it predicts the final outcome with great accuracy compared to many other classification algorithms. As we can see in Figure 2.12, random forest is a collection of multiple decision trees where each decision tree produces their own prediction. The class that is predicted the most, will be the final prediction of the random forest as shown in Figure 2.12. Each decision tree is built using random subset of features from the training data.

For example, lets assume person A wants to pick a place, he would like to visit. He goes to a friend and his friend in return asks some questions such as where he travelled the last time and if he liked it or not. Person A then again goes to another friend and this friend asks whether he likes mountain or sea and less or more tourists. Therefore, the type of questions varies for each person and these questions are the different features that would predict the place he could visit. Thus the common answer

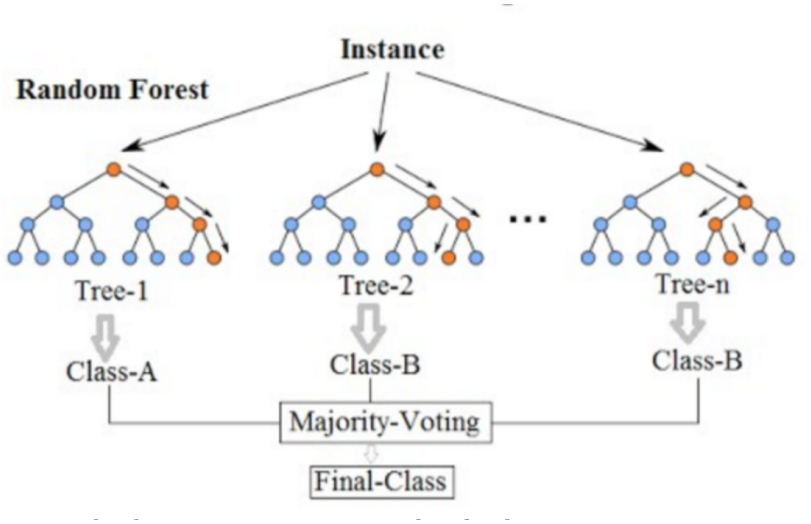


Fig. 2.12 Multiple decision trees to predict the final outcome [44]

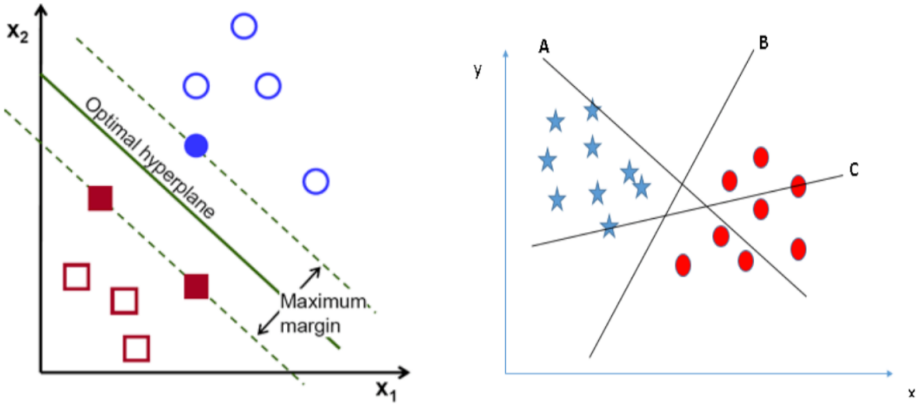
from multiple decision trees are finally recommended.

To improve the speed of the model or to improve the predictive accuracy hyper parameters are usually used which are built in random forest functions. Random forest avoids over fitting. It is also very easy and quick to train a model in random forest however, in real time it will take a lot of time to predict the final outcome. Another great advantage of random forest is that it can be used to extract the important features that are needed for prediction and all the irrelevant features that are not needed can be omitted for prediction thus giving higher accuracy and faster prediction.

- **Support Vector Machine**

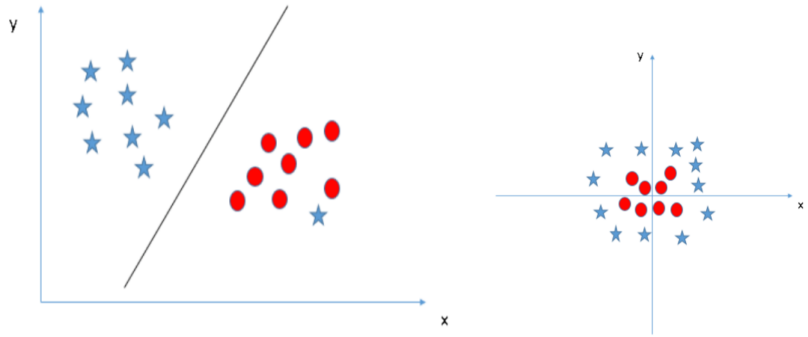
"Support Vector Machine" (SVM) is a supervised machine learning technique that can be used to solve classification and regression problems using a set of given training data [16, 60, 24]. However, most of the time it is used for solving binary classification problems. SVM can handle missing data's. SVM algorithm follows a certain technique to predict the final result. This is where the terms such as hyperplane, margin and kernel comes in. Therefore, the first task in applying the SVM technique is to plot all the data points on an n-dimensional graph where n is the total number of features and the value of each feature will be a specific co-ordinate in the graph. The main task of SVM is to find the optimal hyperplane that will separate the two class from each other which can be seen in Figure 2.13a, Figure 2.14b, Figure 2.14a, Figure 2.15b. It

is called hyperplane since it can work with multi dimension.



(a) (b)

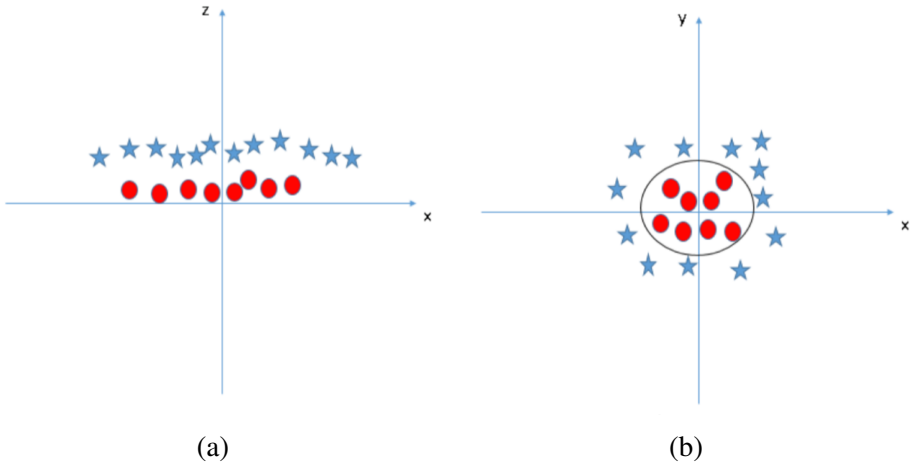
Fig. 2.13 (a)Optimal hyperplane using maximum margin [60]; (b)Separating classes and selecting optimal hyperplane [24]



(a) (b)

Fig. 2.14 (a)Optimal Hyperplane with outliers present [24]; (b)Non linear data points [24]

Now, the question comes how do we choose the optimal hyperplane. In Figure 2.13b, the optimal hyperplane is line B since we can separate the two class in such a way so that most of the points of a single class goes to one side and the points of the other class goes to the other side of the line thus giving the best accuracy. Moreover, it is possible to separate the training data into distinct classes in such a way that there will be several hyperplanes as we can see in Figure 2.13a. In this case, the hyperplane is optimal only when the margin between the training data is maximum. From figure 1 we can see, to calculate the margin we should find the distance between the hyperplane and the closest data point and double it. The maximum margin where there will be no



(a) (b)

Fig. 2.15 (a)Using kernel to convert into linear form [24]; (b)Non linear data as seen in the original input space [24]

data points within the margin will be considered the optimal hyperplane. We select the maximum margin because having a low margin means there is a high chance of miss classification. There maybe some outliers in the training data such as in Figure 2.14b, which are ignored by the SVM algorithm and hence we can still find the optimal hyperplane with the highest margin.

From Figure 2.14b we see, that in some cases, the data points cannot be separated linearly, hence, a new feature of higher dimensional input space is calculated. Thus we introduce an additional feature 'z' where z is equal to sum of squared x and squared y. Then we plot a graph of 'x' against 'z'. In figure 2.15a, we see that the classes can now be separated linearly. This method is know as 'kernel' technique which is used to separate classes which are not linearly separated. This feature is a part of SVM, which is a function that converts low dimensional input space into higher dimensional space after which the hyperplane in the original input space will look like the circle show in the Figure 2.15b.

Data Collection and Processing methods will be discussed in the next chapter in details.

Chapter 3

Data Collection And Pre-processing

A comprehensive description of the process of designing the survey form, data collection and cleaning procedure and a statistical analysis of the data has been presented in this section.

3.1 Recruitment and Procedure

The BRAC University's Computer Science and Engineering Department and BRAC University's counseling unit has approved this research study, it's study procedure and measure. All subjects participated on their own consent and had to sign a consent paper before doing the survey. During this research our main concern was about participant's data privacy. For this we have ensured all the participant that no personal information like name, address, student id, email id and contact number will not be collected from them during the survey.There is no way that we can back track and identify any user from data we have collected. In this way we have protected and maintained our data privacy of our participant taking the survey. The survey was taken from 6th July, 2018 till present and collected data from 935 students. Furthermore we are continuing to collect more data for our further research.

The participant for this study was all private university's undergraduate students across Bangladesh. Majority of our survey was done by BRAC University and North South university undergraduate students and there are other few other universities from where undergraduate students had participated on their own consent . Before collecting the data we have we have gone through extensive meetings with Ms. Anne Anthonia Baroi and Mr. Shami Suhrid, lecturers and members of the counseling unit of BRAC University. They helped through the whole research and taught us the exact procedure how to collect data when dealing with psychological related surveys. We have gone through many survey samples of their previous work and come up with our own survey questionnaires following the standard

structure and requirements. In addition to this, they have taught us how to deal with the participants when performing the surveys for instant we have to respect participant's decision like if they are willing to quit the survey at any moment,they have the right to do so. We have considered all the requirements given by them when dealing with the participants. Therefore, we convinced them to participate in the survey only if they are willing to do so by making a small video. In the video we talked about our motive and what we are planning to achieve from our research work. We also talked about the rules that individuals should follow while taking part in the survey. The rules are as follows:

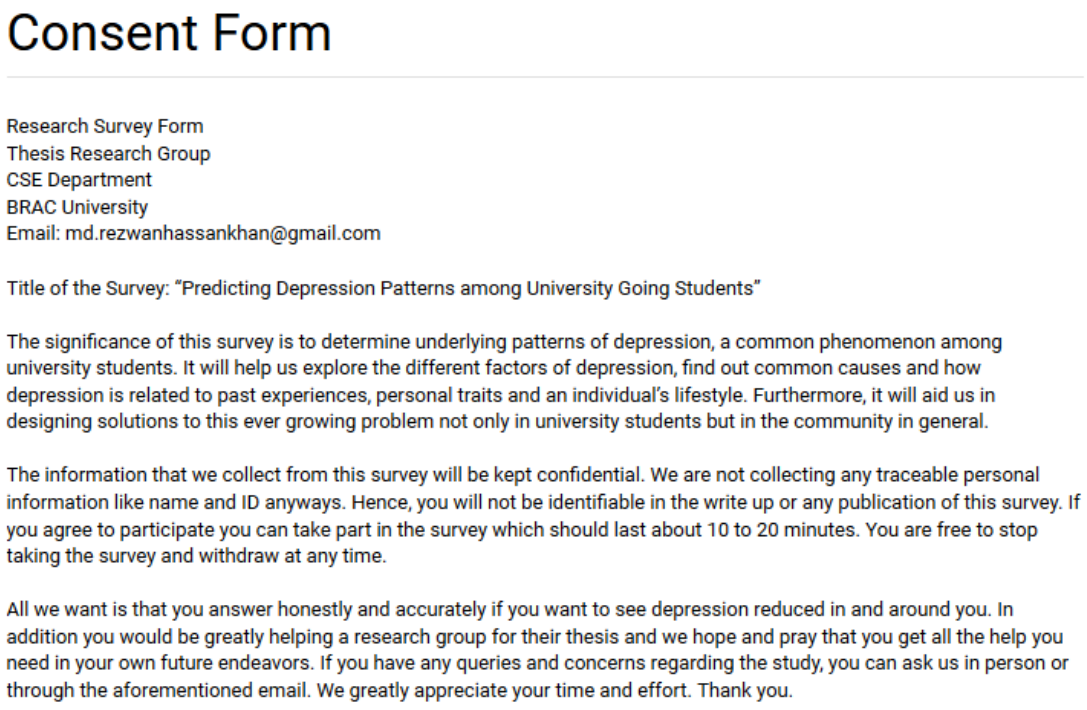
- People taking part in the survey should try their best to complete the survey within 15-20 minutes.
- Students should answer the two scaling systems based on the recent feelings that is at most 1 week.
- They should not think the answer to a question twice. Rather they should select the first option that comes to their mind.
- A participant can leave the survey if they do not want to continue the survey at any given of time.
- Participant has to be truthful about their feelings and must not fill up the survey with incorrect information.

3.2 Data set description

3.2.1 Survey Questionnaire

Our survey paper consists of a set of questionnaires which assess the psychological functioning of the participants. It has four sections, first section consist of the consent form which the user have to agree with to continue with the next section. In addition ,this section contain seven questions to know some basic information (age, sex ,academic year, medium of school, department and cumulative grade point average) about the participant. In the consent form, we gave them a brief idea of what we were doing and that the students are taking part at their own will. We also mentioned that they cannot be tracked back at any cost and if they want they can quit at any moment of the survey.

Second section consist of a questionnaires where there were 16 questions. These 16 questions are the most relevant reasons which lead to depression in a person which we have found out



Consent Form

Research Survey Form
Thesis Research Group
CSE Department
BRAC University
Email: md.rezwanhassankhan@gmail.com

Title of the Survey: "Predicting Depression Patterns among University Going Students"

The significance of this survey is to determine underlying patterns of depression, a common phenomenon among university students. It will help us explore the different factors of depression, find out common causes and how depression is related to past experiences, personal traits and an individual's lifestyle. Furthermore, it will aid us in designing solutions to this ever growing problem not only in university students but in the community in general.

The information that we collect from this survey will be kept confidential. We are not collecting any traceable personal information like name and ID anyways. Hence, you will not be identifiable in the write up or any publication of this survey. If you agree to participate you can take part in the survey which should last about 10 to 20 minutes. You are free to stop taking the survey and withdraw at any time.

All we want is that you answer honestly and accurately if you want to see depression reduced in and around you. In addition you would be greatly helping a research group for their thesis and we hope and pray that you get all the help you need in your own future endeavors. If you have any queries and concerns regarding the study, you can ask us in person or through the aforementioned email. We greatly appreciate your time and effort. Thank you.

Fig. 3.1 Consent form

stated in Chapter 2. Along with this, section 3 and section 4 consists of Beck Depression Inventory-II (BDI-II) scaling method and Depression Anxiety Stress Scales - Bangla Version (DASS 21-BV) scaling method respectively to determine whether a student is depressed or not.

3.2.2 Relevant Causes of Depression

Although the symptoms of depression may be similar across different age groups, the reasons people are diagnosed with depression differ among teenagers, adults and older people in general. This can be understood by inference of course as people live quite different lives during different phases of their life and hence face different problems. Since our research focuses on undergraduate university students, much of our background work on finding out the relevant reasons for depression was done on studies related to depression in younger people. They are described as follows.

According to the author [45], social life, psychological factor and academic life plays a important role which may lead medical students into depression and may trigger to abuse

drug and suicidal behavior. According to the article the high rate of depression among students are for problems in social life like family problem, depression in family members, drug and alcohol addiction. Along with that problems in academic life also may trigger depression like poor academic result and poor performance.

Psychological problem is one of the most important factor that may cause depression.The low income country's university's undergraduate students face with psychological morbidity like financial problem, tension about career, drug and physical abuse, academic pressure and workload which may lead to depression [35]. Along with this excess use of technology, mobile phone and social media may lead to isolation and sleep disturbance thus leading to depression or mental illness related problem [64].Depression is also a genetic problem. So students may also get this problem from their parent if they have depressive symptoms. Cyber-bullying is also another reason which may lead to psychological problems in students.

Along with this we have found out most relevant reasons which may lead to depression through extensive research and taking help from BRAC University counseling unit and they are described below:

1. **Family history of mental illness**

   If there is a history of depression in the family, descendants are at a higher risk of depression. A 30-Year Study of 3 Generations at High Risk and Low Risk for Depression reveals that grandchildren with two previous generations affected by depression was at the highest risk of being diagnosed with depression [77]. Myrna M. Weissman, PhD, of Columbia University and colleagues wrote that the offspring of depressed parents being at a greater risk of psychiatric disorders was established. They sought to find out whether depression beyond two generations was significant in grandchildren. The researchers interviewed a longitudinal retrospective cohort family sample of 251 grandchildren a mean of two times and their biological parents a mean of 4.6 times and grandparents up to 30 years [54]. The findings revealed that the risk for depression in children was significantly higher for parents who had suffered from major depressive disorder (MDD) and the risk increases even further with a history of depression among grandparents.

   Biological relatives with a history of psychiatric illness contribute to the increase in risk of most psychiatric illnesses in individuals [78]. Another research shows that the risk for depression in biological children increases from 12.6% to 41.4% if grandparents or

great-grandparents had depression [58]. Thus we can see that a history of this disorder in the family contributes towards depression in later generations.

2. **Academic condition**

   Varsha Srivastava, a student at Boston University talks about how her poor grades made her anxious and put her in stress until eventually she plunged into a state of depression [75]. Although we have not found enough evidence from literature to say that poor academic performance can lead to depression, the converse that depression can lead to academic failures or a drop in CGPA for university students is well known and proven. Hysenbegasi, Hass, and Rowland (2005) conducted a study on undergraduate students of the Western Michigan University and found out that students diagnosed with depression at the campus Health Center had a GPA of half a letter grade or 0.49 points lower, as compared to a control group who had not been diagnosed with depression [31]. Another research carried out among second year college students in the city of Rawalpindi, Pakistan, showed that depression had a negative effect on the academic performance of the students and that their performance differed considerably depending on their level of depression [43]. Therefore we considered this feature worth including in our study.

3. **Drug addiction**

   Addiction to drugs or substance abuse is strongly correlated with depression according to several studies. On a report of the National Bureau of Economic Research, 69 percent of the America's alcohol and 84 percent of the America's cocaine consumption is done by people who have at least once been diagnosed with depression [71]. The comorbidity of drug abuse with depression is well established [76]. Substance abuse can alter the central nervous systems in ways that can cause depressive symptoms like feeling sad and hopeless in individuals. On the other hand, depressed people turn to drugs to relieve stress and rejuvenate their spirit. Thus the two conditions can persist and worsen each other in a state defined as Dual Diagnosis. One in four people suffering from dual diagnosis are at the risk of committing suicide [65]. According to the Journal of Clinical Psychiatry, one out of three adults who are addicted to drugs also suffer from depression [DualDiagnosis.org]. A research conducted among college students finds out that alcohol and substance abuse is associated with Major Depressive Disorder (MDD) [19]. Hence there is sufficient evidence that there might be a relation between drugs and depression.

4. **Relationship/Affair**

   Depression might affect couple relationships and vice versa. There is proof that people who have relationship problems are three times more prone to depression than people who are not and 60% people suffering from depression consider relationship difficulties to be the main cause [Rel]. On the other hand, the National Institute of Health alerts people that being single increases the risk of depression [Butler]. It is quite natural for people to be troubled due to problems with their partners. However, if irritations, pessimism, restlessness and other symptoms remain for a long time, someone may go into depression. Single people on the contrary, may also face symptoms like loneliness, fatigue etc. and eventually plunge into a state of depression. University of Waterloo psychologist Uzma Rehman and colleagues (2015) say that people who have clinical depression are not content with their relationships [79]. Breakups, too can cause temporary sadness and sleep problems but only if these and other problems persist for over two weeks, can someone be affected by depression. We have considered all these factors in a sequence of questions in our survey form.

5. **Conflict with friends**

   In a way somewhat similar to couple relationships, difficulties in relationship with friends can also lead to depression in individuals. If someone has regular conflict with friends, this will definitely affect their feelings and in some cases make them extremely sad for a long period of time causing depression. Conversely, depressed people tend to have troublesome friendships as they can be extremely demoralizing and difficult to comfort [7]. One research shows that young people who are disagreeable might not face relationship problems, but their relationships are not very good according to their friends and romantic partners [27]. Even though there is not considerable evidence in the literature that associates depression with friendship conflicts, it is nonetheless a variable of consideration.

6. **Financial problems in your family**

   In an increasingly materialistic world, money or the lack of it dominates much of what we feel and do. Hence, it is no surprise that financial problem in the family in turn leading to stress and the feeling of burden on children and less money to spend on enjoyable activities can thrust a university student into depression. Debt in the family can further worsen the situation. A study reveals an alarming 24% rise in depression

symptoms with a 10% increase in short-term debt [52]. While parents will definitely be more affected by such scenarios, the effects on children cannot be ignored. This might also turn into a vicious cycle where poor financial health leads to poor mental health, which leads to increasingly poor financial health, and so on [15]. A research conducted by the University of Southampton and Solent NHS Trust among 454 first year British undergraduate students confirms an increasing risk of mental health problems such as depression and alcohol abuse due to financial problems and worrying about debt at university [63]. Therefore, as can be observed, the link between depression and financial problems in the family is well known.

7. **Violence in family**

   We made it clear to the respondents of the questionnaire beforehand what we mean by violence in the family. Domestic abuse or violence here means any threatening or violent behavior between family members, both physical and emotional and in particular such misconduct between parents that can hamper a child's or a teenager's well being and affect them throughout their lives. Members of the family being violent with each other both in speech and action can cause children to become aggressive themselves and be disobedient escalating furthermore to instances of alcohol and substance abuse [49]. All of these in turn can cause someone to become depressed. Hiremath Debaje (2014) concludes from a study conducted among adolescents in Mumbai, India, that adolescents can suffer from varying degrees of depression from mild to severe as a major effect of domestic violence [29]. They further add that this depression may continue throughout their life resulting in professional and personal shortcomings. Another study says that exposure to a high level of violence significantly increases the risk of major depressive disorder (MDD) [70]. Thus there is a possibility that violence in family can trigger depression in both the short and the long run.

8. **Sadness from death or loss**

   Sadness and grief due to the death of a loved one are common feelings that everyone experience. However, prolonged presence of such a state of grief and hopelessness such that a person's ability to function properly is affected and they begin to prefer isolation can be a sign of depression [49]. Although the American Psychiatric Association has urged doctors not to diagnose major depression in individuals who have recently lost a loved one and specifically listed grief as an exception to the diagnosis of clinical depression, they are considering dropping that exclusion [53]. A research

concludes that the bereavement period is associated with an increased risk of multiple psychiatric disorders and discusses the scope of further studies on the subject [40]. Results from another study tells that losing a parent early in life can cause financial problems in family rendering drug abuse among children and that bereaved groups were more prone to depression and anxiety than non-bereaved groups [37]. They also conclude that while presence of other risk factors beforehand escalates the possibility of psychological and behavioral health problems, even without those factors children who lose loved ones are at risk. Therefore despite the fact that some psychologists' say that depression due to death of a loved one cannot be classified as clinical depression, there is evidence in literature to suggest that the contrary is also a possibility.

9. **Victim of bullying**

   Bullying is trying to exercise power over someone usually someone that the bully or bullies considers weaker than them. It involves hurting an individual physically and emotionally, through pushing, hitting, teasing, offending etc. The fact that bullying has both short term and long term effects on children, teenagers and even adults is well known. Research suggests that children and adolescents who were bullied are more likely to suffer from depression and may even commit suicide [36, 22]. Dr. Andre Sourander, a professor of child psychiatry at the University of Turku in Finland from his research found out that 23% of children who faced regular bullying had some psychiatric problem by the age of 30 [6]. Another study reveals something more interesting and useful in our own research. Wolke Lereya (2015) uncovered that people who were bullied as children are more likely to develop depression that might stay with them and affect them even when they are 40 years old [80]. Hence we have considered past record of bullying as a feature to identify depression in individuals.

10. **Sexual harassment or abuse**

    Sexual harassment in its many forms is generally unwanted touching, gestures, threats and in its most extreme form rape. The victim is usually a female especially in the perspective of Bangladesh. Depression is one of the most common diagnoses of victims of sexual harassment according to Dr. Colleen Cullen, a licensed clinical psychologist [73]. One research findings indicate that sexual harassment is a stressor that is associated with increased depressive symptoms [30]. The researchers also say there is proof that early encounters of sexual harassment in one's career can cause depressive symptoms even when they are adults. This is particularly possible in countries like

Bangladesh where women are urged not to speak about such incidents. The feelings of sadness, guilt, lack of sleep can cause someone to develop depression and trigger thoughts of suicide in extreme cases. Another recent study carried out in 13-17 year olds in UK says that 80 percent teenage girls may suffer from anxiety, depression, post-traumatic stress disorder and other serious conditions four to five months after being assaulted [41]. The reason why is obvious as a sexual harassment itself is a massive social problem that is sadly very much present in every culture and country from the poorest to the richest disrupting the natural way of life and well being of more than half of the world's population.

11. **Hours spent on social media**

Social media is defined as websites and applications that allow people to interact with each other virtually as well as share personal or professional content as a user wishes. The rise of social media usage especially among younger people has been a concern for quite some time. According to the Digital in 2018 report jointly prepared by We Are Social and Hootsuite, 18% of the Bangladesh population, mostly young people use social media actively [62]. The figure stands at a staggering 30 million people, with the most popular platform being Facebook and the device of choice being mobile phones. A research in the US published in the journal of Depression and Anxiety found a link suggesting that most active social media users were 2.7 times more likely to be depressed than least active social media users [9]. Another study carried out in Pakistan tries to find a relation between social media and depression among 200 university students with the help of a questionnaire and Beck Depression Inventory (BDI) [4]. Their results in general showed that more time spent on social media increased chances of developing depression. However, all of these researches suggest that there might be some correlation and a cause and effect statement cannot be made based on these studies. It is one of the most important variables to consider nonetheless given the resounding amount of evidence in the literature.

3.2.3 Participants

In our planning process we have targeted 500 participants to take part in our survey but we were able to break our target and managed to have 935 participants(still counting) which help us immensely to make a very accurate machine learning model. In summer 2018, 7th September, we started our survey and we are continuing to take survey till date. We have taken this survey in three different ways and the time taken complete our survey at max is 25

minutes.

The first was the traditional way where we gave hard copies of our survey form to the participants and they filled it. We have use this method mainly to take one to one survey from individual participant. In the second and third way we have we have used Google survey form where the participant have to fill up the questionnaires in Google form. In second method we have collected data from BRAC University mainly. We have targeted all the labs classes of EEE ,CSE and BBS department. Then have contacted individual department head, the theory teacher and lab teacher of each section prior of our data collection so that they can allocate us 30 minutes from their class timing and took appropriate dates and time from the them . Our fellow batch mates helped us too to take survey from different labs section and we all four members were present in each section when taking the surveys. We were able to cover 1st year lab classes to final year lab classes and the final count of our data collection by the second method was 500 participants in this this way from September 7th till October 30th. The 3rd way we have used is taking online surveys throughout the different universities across Bangladesh and we have started it from September 15th 2018.

As this is a psychological survey we have to take a lot of consideration when taking online surveys as we were not present physically in front of the participants. So to overcome this limitation we have taken a different approach when taking online survey. We have created a short video of 3 minutes which explain what the survey is about.Along with that it states all the rules, requirements and instruction to fill up this form and we have attached our Google form survey description link in that video. After that we have forwarded that video along with our online survey links to all the universities' official Facebook page, university's club pages, different university departments pages and share that video in social media as much as possible. At last we were able to collect around more than 450 data sets and still counting.We have used professional people from BRAC University Cultural Club to make our video as much as appealing to the participant so that they themselves are willing to take part in this survey. To our surprise we were able fulfill our objective and were overwhelmed by seeing the response we get in our online approach. We have found out that people who take part in online survey, they were the most attentive participants among all the participants and the rate of error in their survey form was the minimum compare to the other participants.

We will now look at the statistical data of the survey in the next part.

3.3 Data Cleaning

Finally after collecting 927 students data, we decided to start our prediction based on the collected data. But before running the prediction algorithms on the data, we cleaned the data to remove as many outliers as possible so that we could improve our accuracy. As a result we created many versions of the data. In each version used a certain rule to clean the data's. After cleaning all the data, we had our final version, which is version 9. We were left with 577 data and these data were than used for the prediction model. We will talk about the prediction models in chapter 4.

As we already told that we had 4 parts to this survey. The 3rd and 4th part that is beck scaling system and the bangla scaling system respectively were very important. Since these scaling systems will be used as the basis for our prediction model. We than started cleaning data and saved them into different versions.

Version1: Therefore, to clean the data set, we first looked if there were any missing values in any of the data set. To complete our survey every questions had to be answered to move to the next stage and finally submitting the answer. However, due to some technical problems with google survey system, few of the answers from bangla scaling system for some students were missing. Hence, we deleted all the corresponding rows that had bangla missing answers. We then had 890 rows from 927 rows. We kept the deleted rows in a separate csv file.

Version 2: Edited questionnaires part where some people clicked the radio button with mistake. We had few questions which were to be only answered if the previous question was answered yes. We intentionally kept a radio button instead of a check box to ensure if people were carefully reading all the questions or not. We deleted all the corresponding rows if anyone answered the following questions even after giving no to its previous question. This rows were considered anomaly as these were considered as the people who were not serious enough to take part in our survey. The data set reduced to 800 at this version.

Version 3: We already talked about the two scaling systems. Both of this scaling system had their own individual scores for each of the questions. The scores add up to give a final result to the depression scales. There were multiple classifications based on the final score. However, since we are working on whether a person is depressed or not. We changed normal as "No" and all the other depression classifications as "Yes". Therefore, in this version we assigned the respective scores to each questions based on the standard format of the scaling systems.

Version 4: In this version we added the total score for the bangla scaling system. This bangla score was kept in a newly created column. The respective classification based on the score was then stored in a completely new column named “Bangla Label”.

Version 5: In this version we added the total score for the Beck scaling system. The score for beck was kept in a newly created column. The respective classification based on the score was then stored in a completely new column named “Beck Label”.

Version 6: Before this version most of the mistakes done by students carelessly were removed. In this version we checked the numerical values. We checked if the students entered the correct age since range of the students would usually be from 18-28. Any row with age below or beyond this range was removed. Also we checked the cumulative grade point average. Any row with a value outside the range of 0-4 was removed. However, there were only few cases and thus we finally had 795 rows. Last but not the least we were left with one last check that our psychological department suggested us to do.

Version 7: Basically we used the bangla scaling system as a checking variable for our prediction system. Before this version, we ran our algorithm and we had less accuracy, precision, recall and f-measure. However, after this version our prediction model showed better result and this will be shown in chapter 4. But what we mean by saying that we used bangla scaling system as a checking variable is that we removed all the rows where there were conflicts in the “Beck Label” and “Bangla Label”. If both of the column was not either yes or no, the row was removed. Finally, after this step we were left with 577 data with our final label as either “Yes” or “No” meaning whether a student might be depressed or not.

This was the final data set on which we applied our six prediction algorithms. Based on the final label and our 20 features we ran all the algorithms to see predictions accuracy for different algorithms. We also figured out the relevant features out of the 20 features used for prediction.

3.4 Data Visualization of the Final Data

Data visualization is very important since it helps people to understand the importance of data through visual representation. Sometimes, we might miss some important information such as trends and patterns in the data if we represent it in a text based system. Also important

co-relations between data might be discovered through visual representation. As a result, representing data through visual context is a great way to understand data a lot better. There are many techniques to visualize data. Therefore, for our data we will be using histogram, which is technique to represent the distribution of numerical data. In histogram we represent the total number for bar in the y-axis where each bar represents a specific information written in x-axis. Therefore, we will represent every feature in a histogram and try to understand the data. Here red bar represents “Not Depressed” and blue bar represents “Depressed”. We will see the number of students depressed and not depressed for every answer possible.

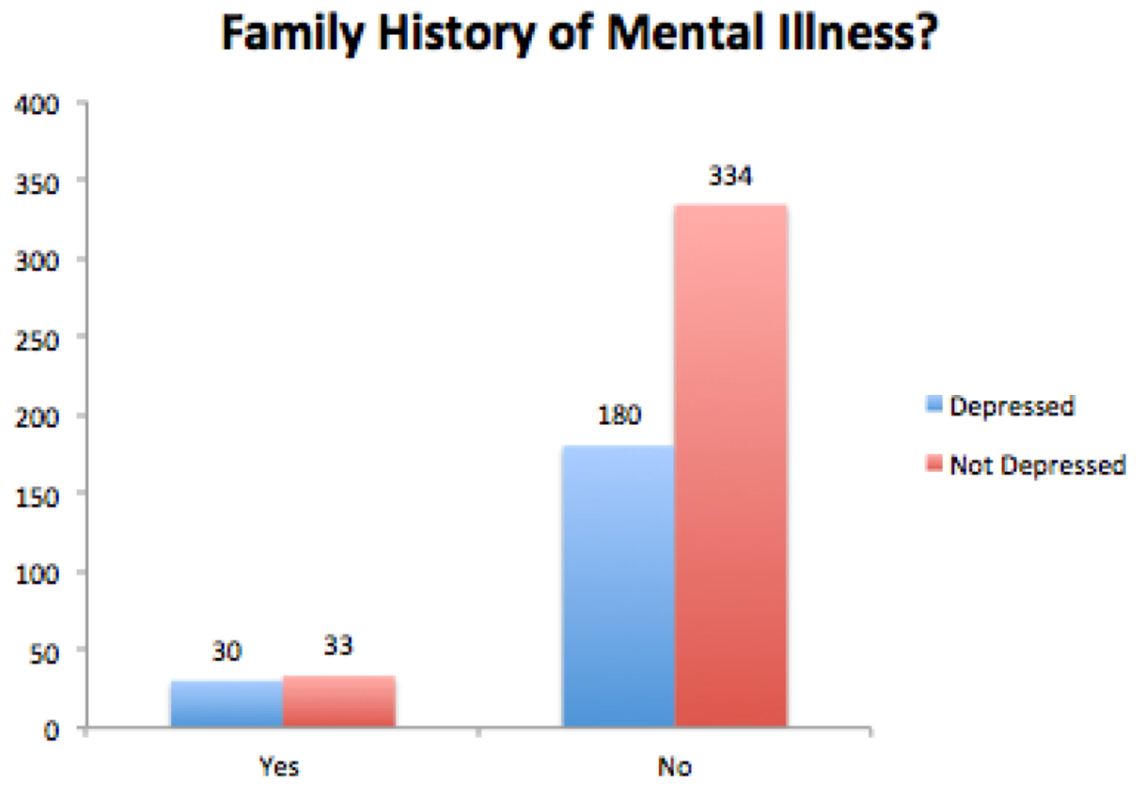


Fig. 3.2 Data Comparison for Family History of Mental Illness

From Figure 3.2 we can see we have more number of students who said they do not have any history of family mental illness. Out of 577 students, 514 said no among which 180 is depressed and 334 are not depressed. Therefore, it can partially be assumed that people who have no problem at home might be less depressed compared to the answer of yes. From yes, we can see that around 50% that is 30 out of 63 is depressed if there is history of mental illness in the family.

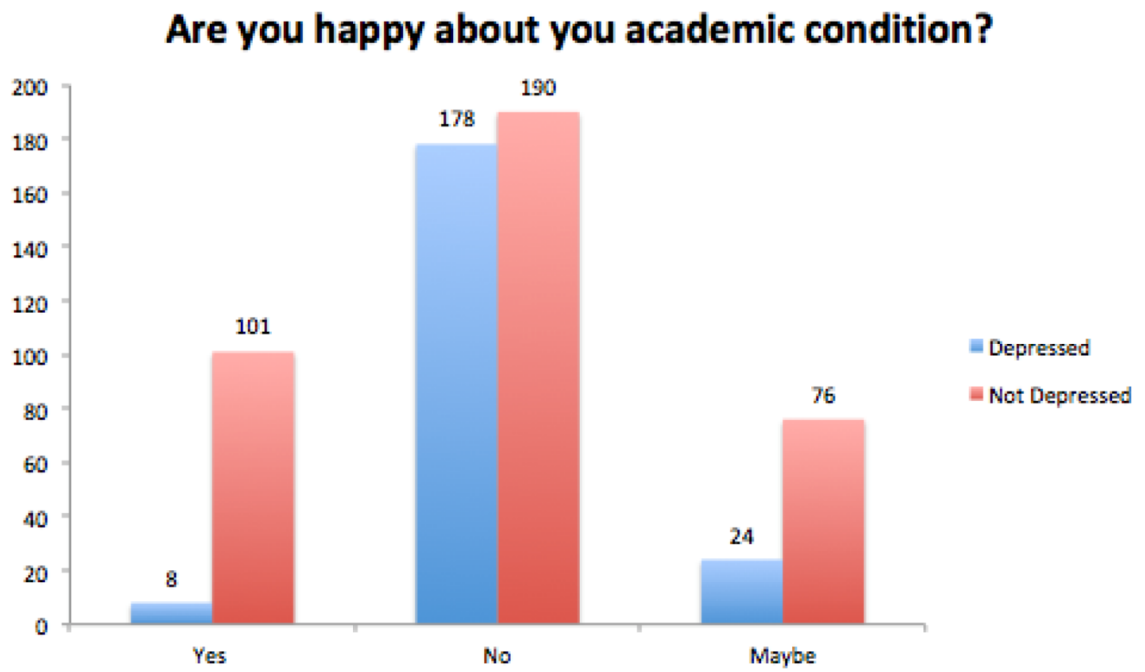


Fig. 3.3 Data Comparison for Are you happy about your academic condition?

From Figure 3.3, we can again see academic condition which includes results, pressure from the family about the grades, huge study load and understanding the regular classes. From the visual representation we can see that majority of the people answering yes to this question is not depressed. Only 8 out of 109 is depressed whereas 101 student is not depressed. Also, students who said "no" to whether they are happy about the academic result shows how strongly the feature influences depression among students. About 50% that is 178 out of 368 is depressed when they say no to this question. Therefore, we can say that this question is one of the major factors that helps in predicting depression.

From Figure 3.4, it is tough to come to a conclusion by just seeing the visual representation of the following data. This is because we have very less information for the answer of "Yes" to this question. We only have 29 yes out of which 8 is depressed and 21 is not depressed. But we can at least assume that if someone is not taking drugs there is a chance that he will have less chance of depression. There are 548 students who are not taking drugs out of which 202 are depressed and 346 students are not depressed.

Question in Figure 3.5a and Figure 3.5b is correlated. The question to Figure 3.5b was only answered by the students if he/she selected "Yes" to the question in Figure 3.5a. From Figure 3.5b we can see that 229 out of 359 students are not depressed even though they are

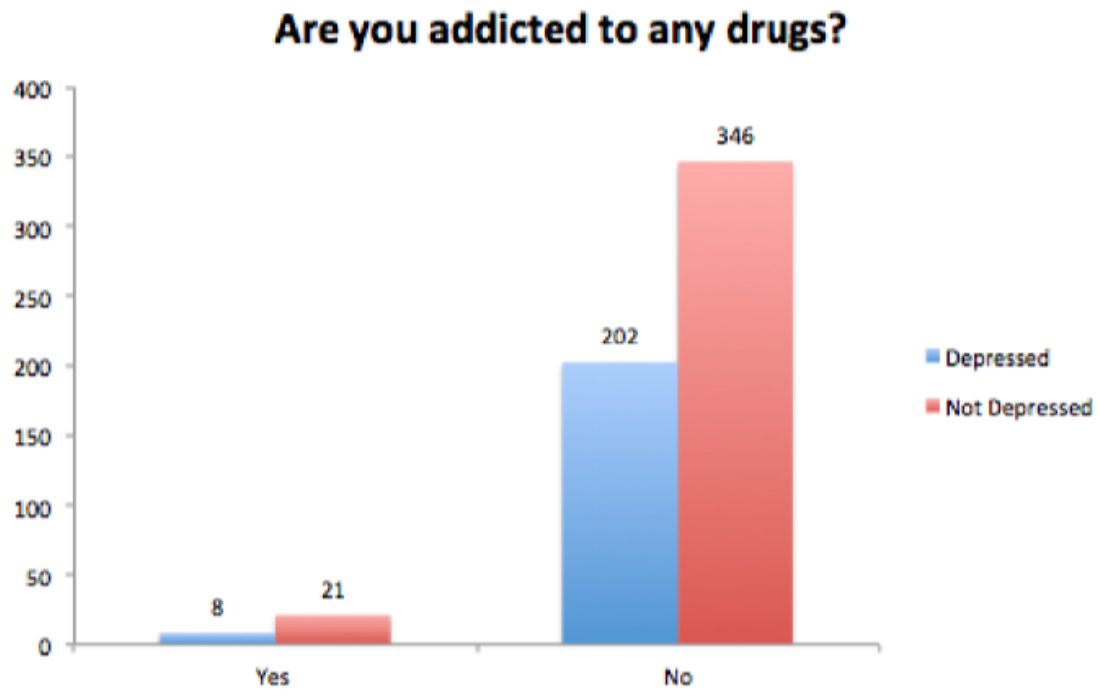
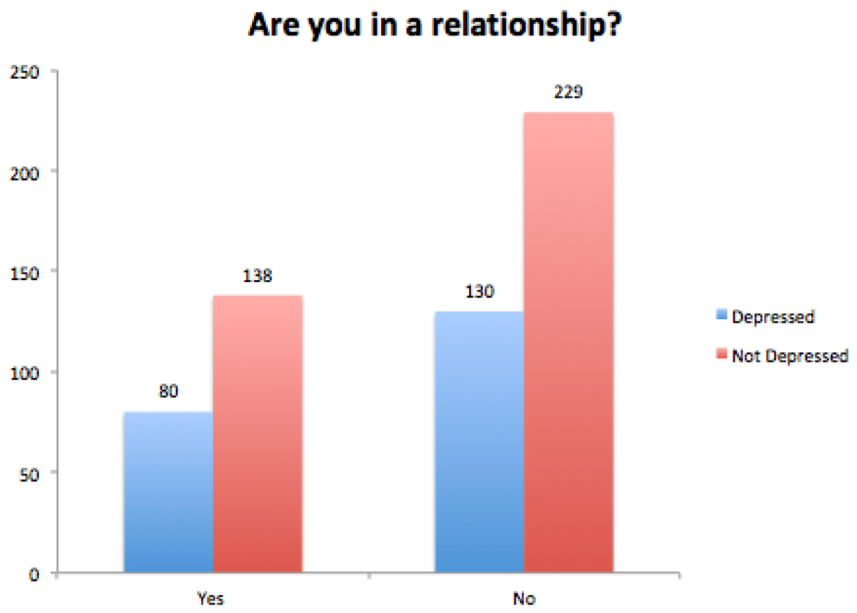


Fig. 3.4 Data Comparison for Are you addicted to any drugs?

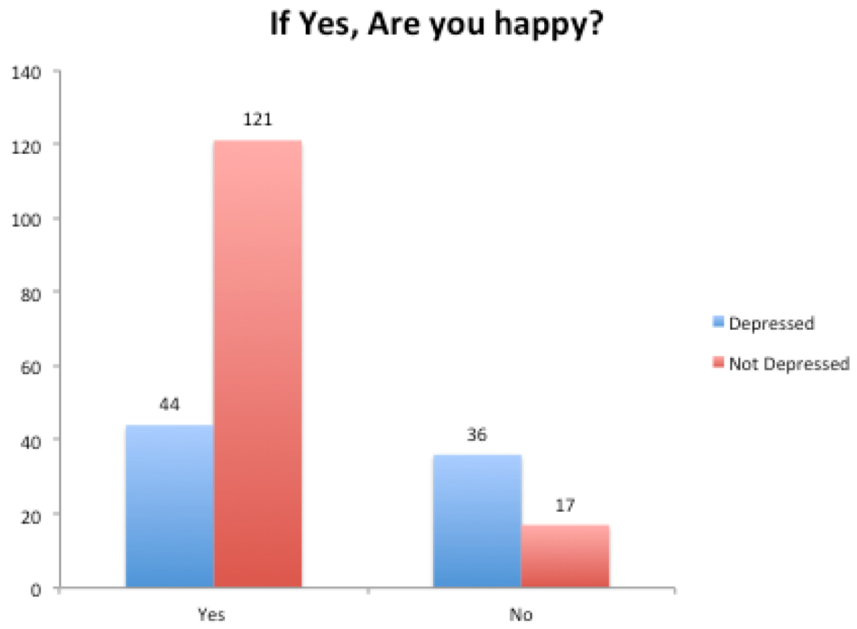
not in a relationship. Therefore having a relationship might not affect the final outcome but whether a person is happy or not with the relationship might affect a student. We can see from Figure 3.5b, that a student who answered "yes" that he/she is happy in their relationship shows that 121 out of 165 is not depressed. This tells majority of the students are not depressed if they are happy with their relationship. However, from students saying "no" we can see majority of is depressed in their life. 36 students out of 53 students are depressed when they said no to the question in Figure 3.5b.

Figure 3.6 shows whether students are depressed or not if they had a recent breakup. We can see more than 50% that is 49 out of 96 students are depressed if they go through a recent breakup. We can also see that 320 students out 481 are not depressed if they do not go through a break up. Therefore, we can assume that question in Figure 3.6 plays a vital role in predicting depression.

From Figure 3.7 we can conclude that students who are getting involved in conflicts with their friends are more depressed in their life than students who do not get into fights with their friends. We can see that students answering "Most of the time", 10 out of 15 and "Often", 29 out of 59 have more cases of depression than students answering "Rarely" and "Never". For



(a)



(b)

Fig. 3.5 Data Comparison for (a) Are you in a relationship?; (b) If Yes, Are you happy?

students answering rarely there are 278 out 425 and for never 54 out of 78 are not depressed. Therefore majority of them are not depressed when they do not go into conflicts. Hence, we can conclude that question in Figure 3.7 co-related to our prediction.

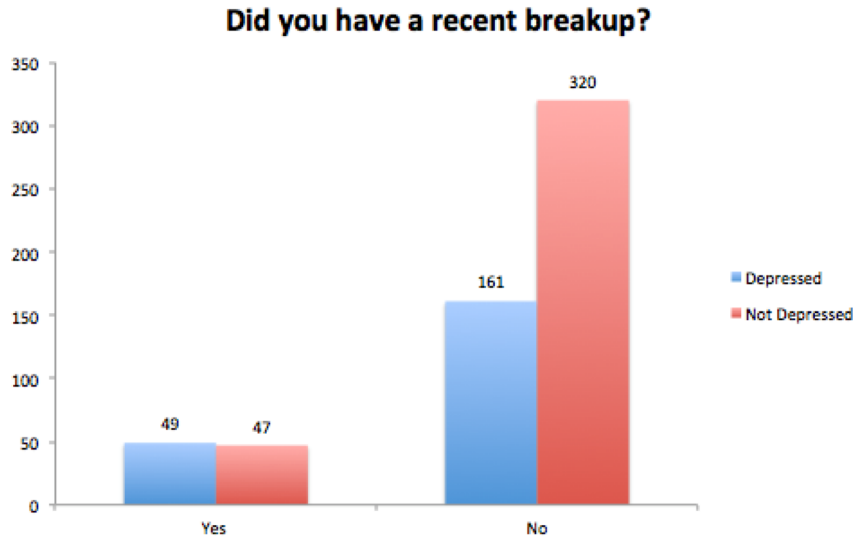


Fig. 3.6 Data Comparison for Did you have a recent breakup?

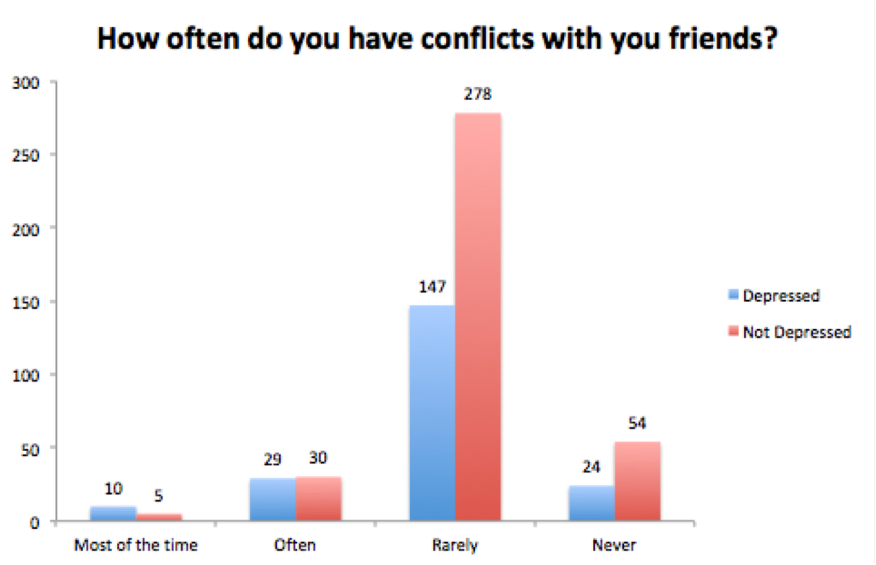
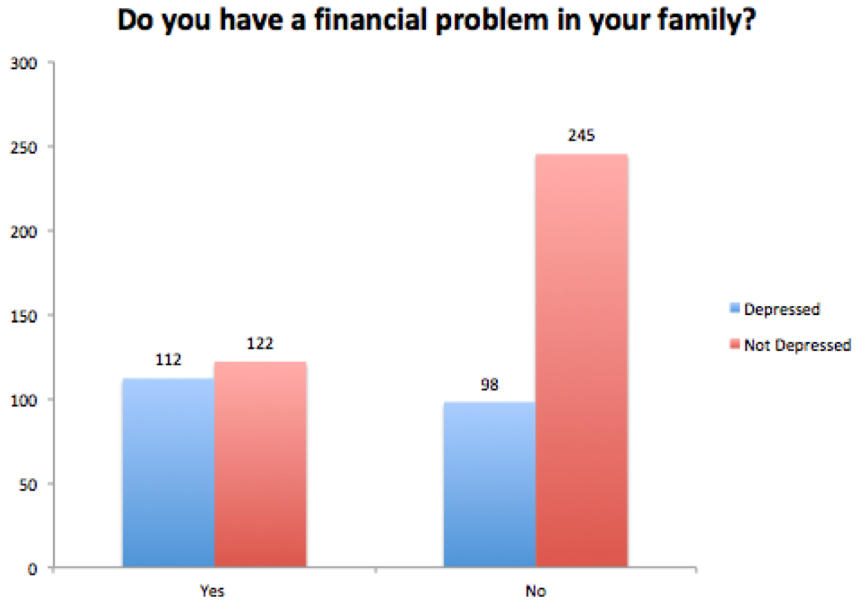
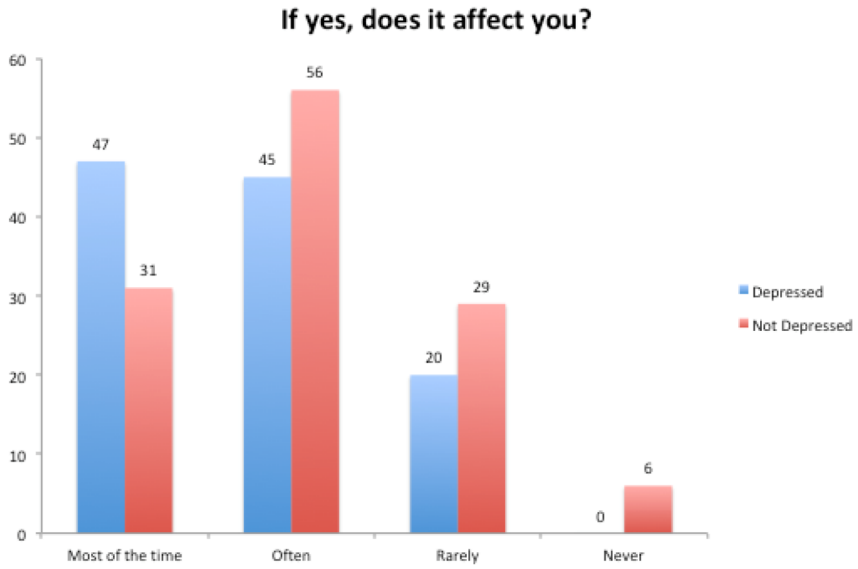


Fig. 3.7 Data Comparison for How often do you conflict with your friend?

Question in Figure 3.8a influences depression a lot. We can conclude that by seeing the "Yes" and "No" to the following question. Students who said yes to financial problem in family are more or less depressed. Around 50% that is 112 out of 234 is depressed for the



(a)



(b)

Fig. 3.8 Data Comparison for (a) Do you have financial problem in your family?; (b) If Yes,does it affect you?

case. However, students who do not have any financial problem are less prone to being depressed which can be concluded from Figure 3.8a. There 245 out of 343 students who are not depressed when they do not have any financial problem. Figure 3.8b is again related to financial problem. Students who answered “yes” to financial problem only answered the

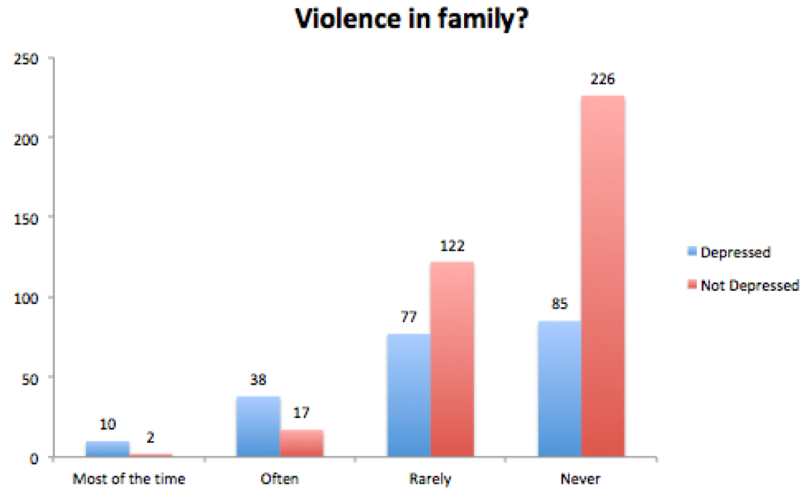


Fig. 3.9 Data Comparison for Violence in family?

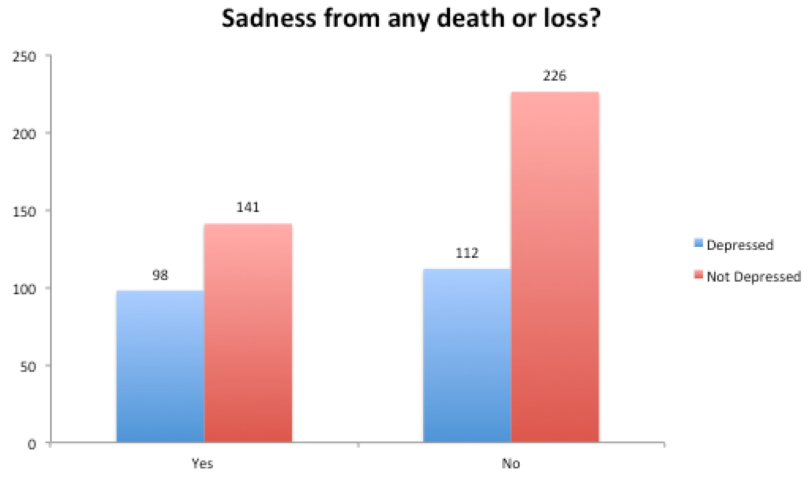
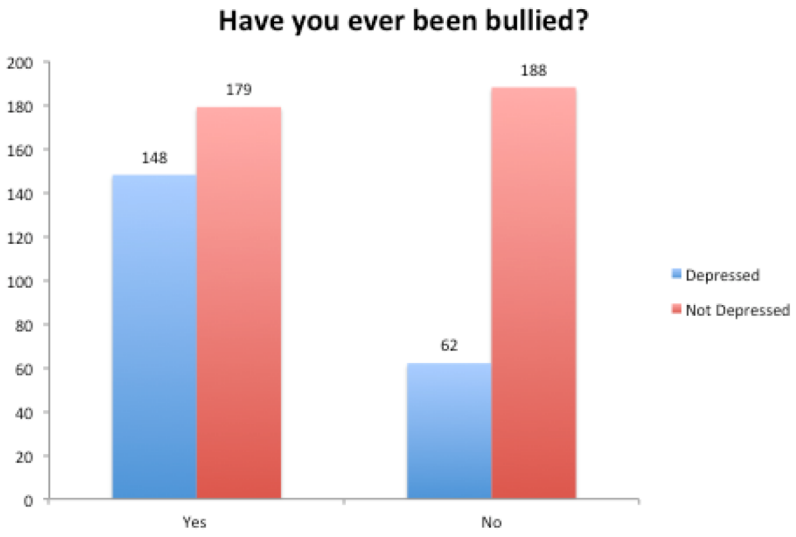


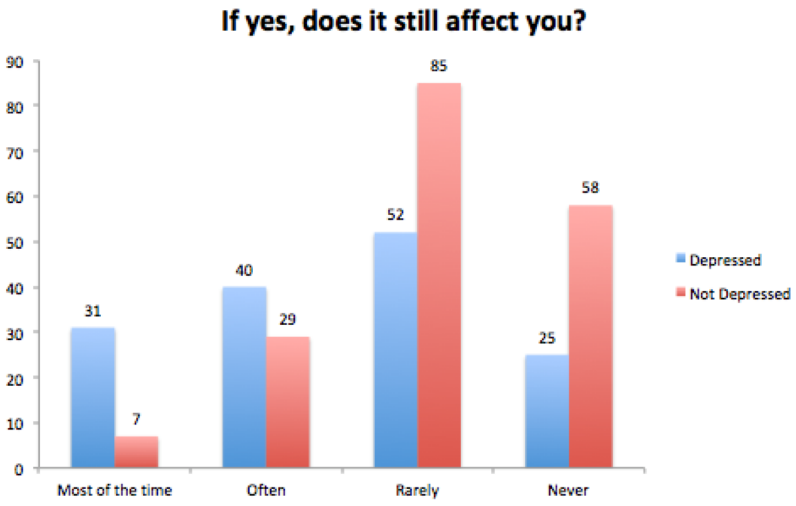
Fig. 3.10 Data Comparison for Sadness from any death or loss?

following question whether it affects them or not. We can clearly see that financial problem affects students greatly by seeing the answer to the question of whether it affects them or not. Students answering “Most of the time” and “Often” are more depressed that is 92 out of 179 is depressed due to financial troubles whereas not a single person is depressed if it never bothers them.

Violence in family includes members of the family being aggressive physically and verbally. We can also see how important this question is for prediction of depression from Figure 3.9. 48 out of 67 students are depressed when there is either a violence in the family most of the time or often. This is way more than 50%. Also students who rarely or never



(a)

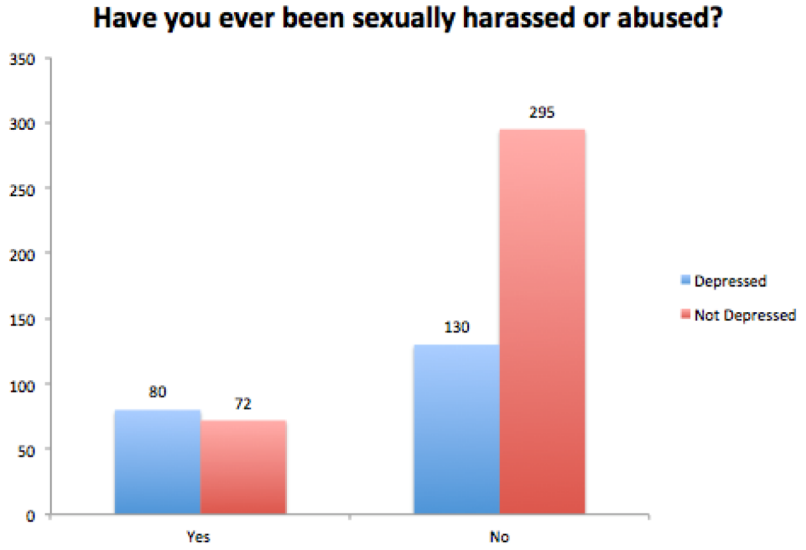


(b)

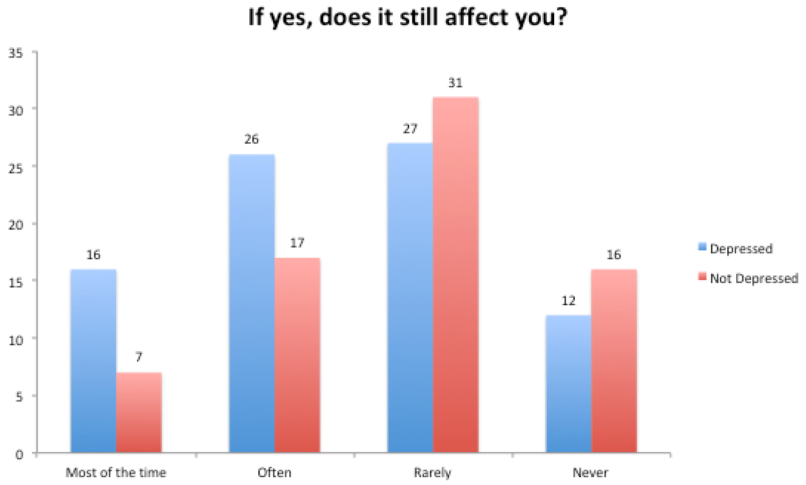
Fig. 3.11 Data Comparison for (a) Have you ever been bullied?;(b) If Yes, does it still affect you?

have violence in their family are more inclined towards being not depressed. 348 out of 510 are not depressed since they do not have any family violence. Therefore, we can again assume this feature as a very important factor to depression prediction.

From Figure 3.10 we cannot just assume whether a student is affected by the death or loss of their closed ones. Because for both “Yes” and “No” the answer for not depressed is more. Less than 50% that is 98 out 239 is depressed when they say yes to the following



(a)



(b)

Fig. 3.12 Data Comparison for (a) Have you ever been sexually harassed or abused?; (b) If Yes, does it still affect you?

question. Hence we cannot just say whether this question influences our prediction or not by just seeing the visualization.

Again Figure 3.11a and Figure 3.11b is correlated. Students only answered to the question in Figure 3.11b only if they answered to question of Figure 3.11a. The question was if they have been bullied or not. Just by seeing the pattern in this question we cannot assume whether this question affects the final outcome or not. For both “Yes” and “No” we see that majority of

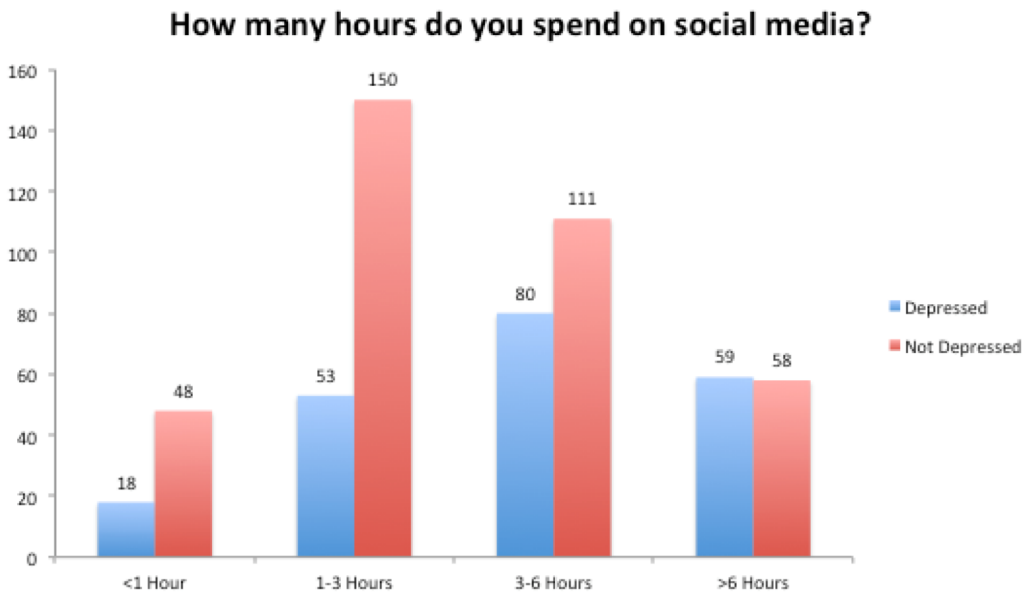


Fig. 3.13 Data Comparison for How many hours do you spend on social media?

the people are not depressed. For Yes, 148 out of 327 is depressed which is less than 50%. However, we can only understand the importance of this question by seeing the second part as shown in Figure 3.11b.

Student who are still affected by bullying is more depressed than students who are not affected. From Figure 3.11b we can see that students who are affected "most of the time" and "often" are more depressed. 31 out of 38 which is more than 80% students are depressed when it affects them most of the time and 40 out of 69 is affected when it affects often. Although there are students who answered rarely and never and still have cases of depression but the majority of the people is not depressed. Hence we can again conclude, if a person is bullied and it affects them, than there is a greater chance that he/she has depression.

Again Figure 3.12a and Figure 3.12b is related to each other. Students only answered to the question in Figure 3.12b only if they answered to question of Figure 3.12a. The question was if they have ever been sexually harassed or abused. Here, the first question that is shown in Figure 3.12a influence depression in a student. We can assume by seeing the number of students depressed when they have been abused or harassed. Again more than 50% which is 80 to 72 ratio are depressed and 295 to 130 ratio not depressed when they are not harassed. We can further justify to our assumption by seeing Figure 3.12b. For both "most of the time" and "often", depression is influenced as we can see from the figure.

The representation shows 42 out of 66 are depressed for often and most of the time. This is more than 50%. Also for “rarely” and “never” lots of cases of depression but majority is not depressed. Therefore both of them important factors in our depression prediction.

From Figure 3.13 we can assume that spending time in social media forces student into depression. As we can see the number of hours increase the percentage of depression increases. This is because people who do not have a social life and wastes time on social medias for finding friends are more lonely and depressed than people who spend less time on internet and goes out to enjoy their life. We can clearly see that in Figure 3.13. Only around 25% of student spending less than 1 hour and 25% of student spending 1-3 hours are depressed. This increases to 40% of student depressed when they spend 3-6. This further increases to more than 55% when they spend greater than 6 hours of time on the social media. This greatly shows the importance level of time spending on social media.

From the following section we visualized the data and we assumed some of the reasons are not important factor in our depression prediction. However, this was only an assumption, so in the next chapters we will be using different algorithm to predict depression based on the 20 features and also prediction after optimal feature extraction.

Chapter 4

Experimental Result and Analysis

In the previous chapters we discussed about our proposed model. We talked about how we collected data for our system, how the data was cleaned, and we visualized the final data set using histograms. After that we used different algorithms to detect predictions in the university students. We will shortly be talking about the result for different algorithms. At first, we used our 20 general features that we extracted by reading different articles and ran different machine learning algorithms to find out the accuracy of our model. Again we used recursive feature elimination with cross validation and random forest classification to find out the optimal features that shows that some features in our data set are co-related and we can omit those features for our prediction system [18].

4.1 Applying Algorithms

Our research work aims to find out if a student is depressed or not. Since our data set is based on binary classification where there can any one of the two outcomes. We also used histograms to visualize our data. Since our data gives binary classification and we have few missing values for some of the questions, we used few of the most widely used algorithms for both of the case mentioned. We used 6 algorithms to find out the accuracy, precision, recall, specificity and f-measure of our predictive model. Higher the accuracy and f-measure, the better will be the system. Precision is also an important factor, where precision means the total number of true positive in all the prediction of yes. For our system, this means that a student who is actually depressed from all the predicted depressed results. Recall another important part of prediction system, which is the number of true positives in actual yes results.For our system, this means that a student who is actually depressed from all the actual depressed students. Next, We will compare the accuracy and f-measure for all the algorithms and try to find out the algorithm which will best suit our model. Two more important terms

that are essential for our system, False Negative and False Positive. Lower the False Negative and False Positive, better the model will be. Also, a model will be considered good if both of the are lower and there is a good balance in them. Since our system is predicting depression, it is very important to get lower False Negative as False negative for our model means the system will identify a depressed person as not depressed and False Positive will identify a person who is not depressed as depressed Furthermore, for our system we will try to get at least better False Negative than False Positive as it is very important to at least identify a depressed person as depressed from someone who is not depressed but being identified as depressed. We already discussed about the algorithms in Chapter 2. For training and testing, we used k fold cross validation.

4.2 K fold cross validation

K-Fold cross validation is used to improve the accuracy of the machine learning model.The problem with test/train split lead to over fitting. The number of test set is low compare to train data set which may lead to over fitting. It divides the whole data set in k folds and and each fold will contain the same amount of data in it. One fold is selected as test set and k-1 folds are selected as training set and accuracy of the function is carried out. It is then repeated k times so that every portion of data selected as test set and training set . As we repeated it k times we get k times mean square error. So the the error of this model is computed by taking average of the mean square error over k folds [42, 81].

It is experimentally found out that setting fold value to 10 gives result with low biasing. Along with this it reduces the computation time as it is only iterated 10 times .Every data point is tested exactly once and trained k-1 times [66, 42]. Therefore, we used 10 fold cross validation for each of the 6 algorithms.

4.3 Accuracy for different algorithms with 20 features

We first showed the prediction using all the algorithms with our initial 20 general features.

4.3.1 Deep Learning

From Table 4.1 we can see, that our accuracy is 71.74% which approximately 71%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 132 individuals who were predicted as yes and they are

accuracy:71.74% +/- 6.07(micro average: 71.75%)			
	true No	true Yes	class precision
pred. NO	282	78	78.33%
pred. Yes	85	132	60.83%
class recall	76.84%	62.86%	

Table 4.1 Accuracy, Precision and Recall Of Deep Learning for 20 Features

actually depressed which is the true positive and total number of prediction of yes is 217. Therefore dividing 132 by 217 and then multiplying it with 100, we get a precision of 60.83% approximately 61%. Then we look at the recall, which is the number of true positives in actual yes results. There are 132 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 132 by 210 and then multiplying it with 100, we get a recall of 62.86% approximately 63%. Also, We see a False Negative of 78 out of actual 210 depressed cases which is comparatively lower for our system. False Positive is 85 out of actual 367 not depressed cases.Therefore, a good balance in False Positive and False Negative.

4.3.2 Generalized Linear Model

From Table 4.2 we can see, that our accuracy is 74.17% which approximately 74%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 110 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 159. Therefore dividing 110 by 159 and then multiplying it with 100, we get a precision of 69.18% approximately 69%. Then we look at the recall, which is the number of true positives in actual yes results. There are 110 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 110 by 210 and then multiplying it with 100, we get a recall of 52.38% approximately 52%. Also, We see a False Negative of 100 out of actual 210 depressed cases which is comparatively moderate for our system. False Positive is 49 out of actual 367 not depressed cases.

4.3.3 Gradient Boosted Algorithm

From Table 4.3 we can see, that our accuracy is 71.91% which approximately 72%. Also if we look at the precision, where precision means the total number of true positive in all

accuracy:74.17% +/- 4.49%(micro average: 74.18%)			
	true No	true Yes	class precision
pred. NO	318	100	76.08%
pred. Yes	49	119	69.18%
class recall	86.65%	52.38%	

Table 4.2 Accuracy, Precision and Recall Of Generalized Linear Model for 20 Features

accuracy:71.91% +/- 5.57%(micro average: 71.92%)			
	true No	true Yes	class precision
pred. NO	284	79	78.24%
pred. Yes	83	131	61.21%
class recall	77.38%	62.38%	

Table 4.3 Accuracy, Precision and Recall Of Gradient Boosted Algorithm for 20 Features

the prediction of yes. There are 131 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 214. Therefore dividing 131 by 214 and then multiplying it with 100, we get a precision of 61.21% approximately 61%. Then we look at the recall, which is the number of true positives in actual yes results. There are 131 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 131 by 210 and then multiplying it with 100, we get a recall of 62.38% approximately 62% .We see a False Negative of 79 out of actual 210 depressed cases which is comparatively lower for our system. False Positive is 83 out of actual 367 not depressed cases. Therefore, a good balance in False Positive and False Negative.

4.3.4 K-Nearest Neighbor

From Table 4.4 we can see, that our accuracy is 67.24% which approximately 67%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 67 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 113. Therefore dividing 67 by 113 and then multiplying it with 100, we get a precision of 59.29% approximately 59%. Then we look at the recall, which is the number of true positives in actual yes results. There are 67 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 67 by 210 and then multiplying it with 100, we get a recall of 31.90% approximately 32%.

accuracy:67.24% +/- 4.23%(micro average: 67.24%)			
	true No	true Yes	class precision
pred. NO	321	143	69.18%
pred. Yes	46	67	59.29%
class recall	87.47%	31.90%	

Table 4.4 Accuracy, Precision and Recall Of K-Nearest Neighbor Algorithm for 20 Features

accuracy:74.52% +/- 3.73%(micro average: 74.52%)			
	true No	true Yes	class precision
pred. NO	319	99	76.32%
pred. Yes	48	111	69.81%
class recall	86.92%	52.86%	

Table 4.5 Accuracy, Precision and Recall Of Random Forest for 20 Features

Also, We see a False Negative of 143 out of actual 210 depressed cases which is comparatively a very high for our system. False Positive is 46 out of actual 367 not depressed cases.

4.3.5 Random Forest

From Table 4.5 we can see, that our accuracy is 74.52% which approximately 75%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 111 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 159. Therefore dividing 111 by 159 and then multiplying it with 100, we get a precision of 69.81% approximately 70%. Then we look at the recall, which is the number of true positives in actual yes results. There are 111 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 111 by 210 and then multiplying it with 100, we get a recall of 52.86% approximately 53%. Also, We see a False Negative of 99 out of actual 210 depressed cases which is comparatively moderate for our system. False Positive is 48 out of actual 367 not depressed cases.

4.3.6 Support Vector Machine

From Table 4.6 we can see, that our accuracy is 73.49% which approximately 73%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 103 individuals who were predicted as yes and they are

accuracy:73.49% +/- 5.03%(micro average: 73.48%)			
	true No	true Yes	class precision
pred. NO	321	107	75.00%
pred. Yes	46	103	69.13%
class recall	87.47%	49.05%	

Table 4.6 Accuracy, Precision and Recall Of Support Vector Machine for 20 Features

actually depressed which is the true positive and total number of prediction of yes is 149. Therefore dividing 103 by 149 and then multiplying it with 100, we get a precision of 69.13% approximately 69%. Then we look at the recall, which is the number of true positives in actual yes results. There are 103 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 103 by 210 and then multiplying it with 100, we get a recall of 49.05% approximately 49%. Also, We see a False Negative of 107 out of actual 210 depressed cases which is comparatively moderate for our system. False Positive is 46 out of actual 367 not depressed cases.

4.4 Comparison between the algorithms for 20 features

We will now compare accuracy, precision, recall and f-measure along with the False Negative and False Positive of all the algorithms that we applied. Table 4.1 shows the comparison between all the algorithms.

From Figure 4.1 we can see that except K-NN all the algorithms gave a very close accuracy. Here, blue bar represents accuracy, red bar represents precision, recall is represented using green bar and f-measure is represented using purple bar. From the histogram we can see that we have received the highest accuracy, which is 75% in Random Forest and the lowest accuracy is in K-NN, which is 67%. Even though it seems 67% is not making much of a difference but actually K Nearest Neighbor is not good for our system and this clearly indicated by the f-measure value, which is only 41%. F-measure is more accurate and it measure's the test's accuracy and is calculated using weighted harmonic mean of precision and recall. The higher the f-measure, the better our system will be. Thus K-NN is a bad choice due to less f-measure value. The reason behind this is because, K-NN works best with fewer features when classifying. However, when we train our model, we convert this 20 features using dummy variable finally giving 45 features, which is too much for a K-NN

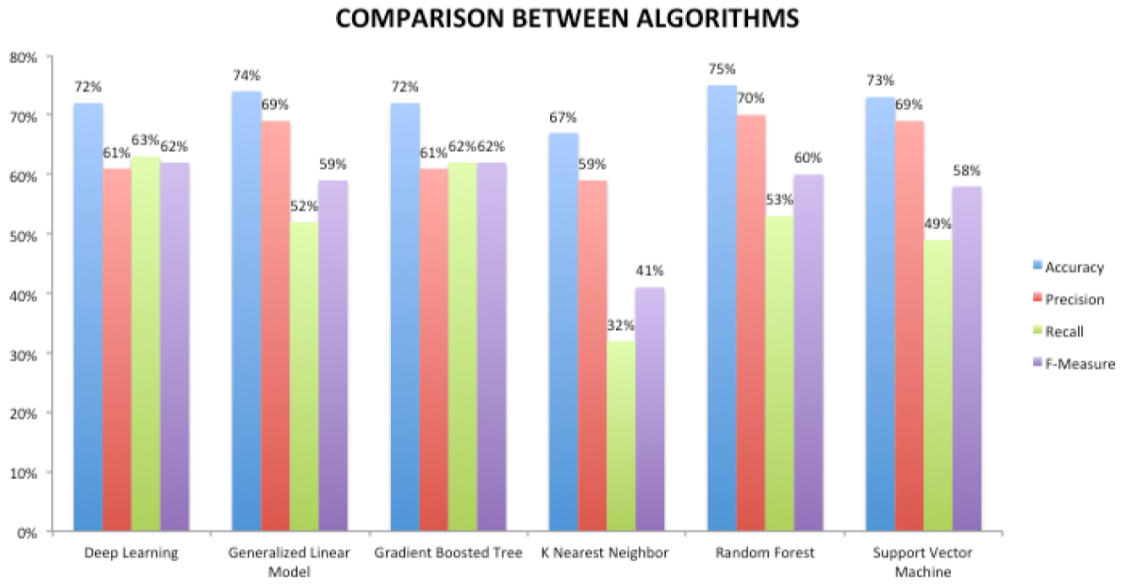


Fig. 4.1 Comparison Between Algorithms For 20 Features

algorithm because in higher dimensions the points are too close to each other and hence difficult to assume from the graph. Also, there is a huge difference in the precision and recall. Precision is still 59% which is not bad enough whereas recall is only 32% which shows that only 32% of the actual depressed case has been correctly identified.

All the other algorithms are also giving good results. SVM has accuracy of 73% with a f-measure of 58%. Generalized Linear Model is giving an accuracy of 74% and a f-measure value of 59%. Both Deep Learning and Gradient Boosted Tree is giving a accuracy of 72% and f-measure of 62%. From the given accuracy and f-measure we can see all of them are more or less close to each other except K-NN. Therefore, K-NN is not at all a good idea for our system.

Moreover, we can see that there is a good balance in False Positive and False Negative and these values are comparatively lower for the algorithms in Deep Learning and Gradient Boosted Tree. Also, from the prediction, accuracy and f-measure, we can come to a conclusion that Deep Learning and Gradient Boosted Tree gives the best result for our model. This can be clearly seen as both of them have the highest f-measure of 62%, which is the weighted average of precision and recall. We can see that precision for both of this algorithm is 61% and recall is 63% for Deep Learning and 62% for Gradient Boosted Tree. There is very less difference in precision and recall for both of this algorithm whereas for other algorithms there is a good difference in precision and recall. 69% precision and 52% recall for Generalized

Linear Model. 70% precision and 53% recall for Random Forest. 69% precision and 49% recall for Support Vector Machine. Therefore, we can conclude that Deep Learning and Gradient Boosted Tree is better compared to others for our system.

We ran different algorithms based on the initial 20 features that we figured out as more important to classify whether a student is depressed or not. After running different algorithms on the 20 features we discussed about the accuracy, precision, recall and f-measure in sub section 4.2 and 4.3. We then found out optimal features using RFE known as Recursive Feature Elimination with Cross Validation and Random Forest Classification. There are different techniques for optimal feature selection but we used this method as it is widely used. This technique helps to reduce over fitting of data, improves accuracy by removing misleading data. It also reduces training time as less features to train the model.

Every time it will give different optimal combination of features, which will be considered as the optimal features for the system. The optimal feature will vary for different data set. For our data set majority of the time the optimal features were close to 15 and thus we took the 15 optimal features that were given by the RFE cross. If we look at Figure 4.2, we can see the change of cross validation score due to change in number of optimal feature selections. We can see that for our data set 15 of the features can be used instead of 20 to get better results. Higher the cross validation score better will be the accuracy. Therefore using this method we also got the 15 optimal features which are more important as shown in Figure 4.3. Therefore, from Figure 4.3 we can say that “F-5”, “F-7”, “F-8”, “F-15” and “F-18” are not required for our prediction. Each of them represents a feature. Here:

F-5: Family history of mental illness?
F-7: Are you addicted to any drugs?
F-8: Are you in a relationship?
F-15: Sadness from any death or loss?
F-18: Have you ever been harassed or not?

We already talked about this in Chapter 3 data visualization part. From there, we almost predicted that these features were not important enough for our prediction model. Hence, our concluding remarks in Data Visualization part are somewhat partially right.

Now, after figuring out the features that would be optimal for our system, we deleted the features and again ran the 6 algorithms and found out the accuracy, precision, recall and

f-measure. Lets see if there is any change in our result and this will be further discussed in the next subsection.

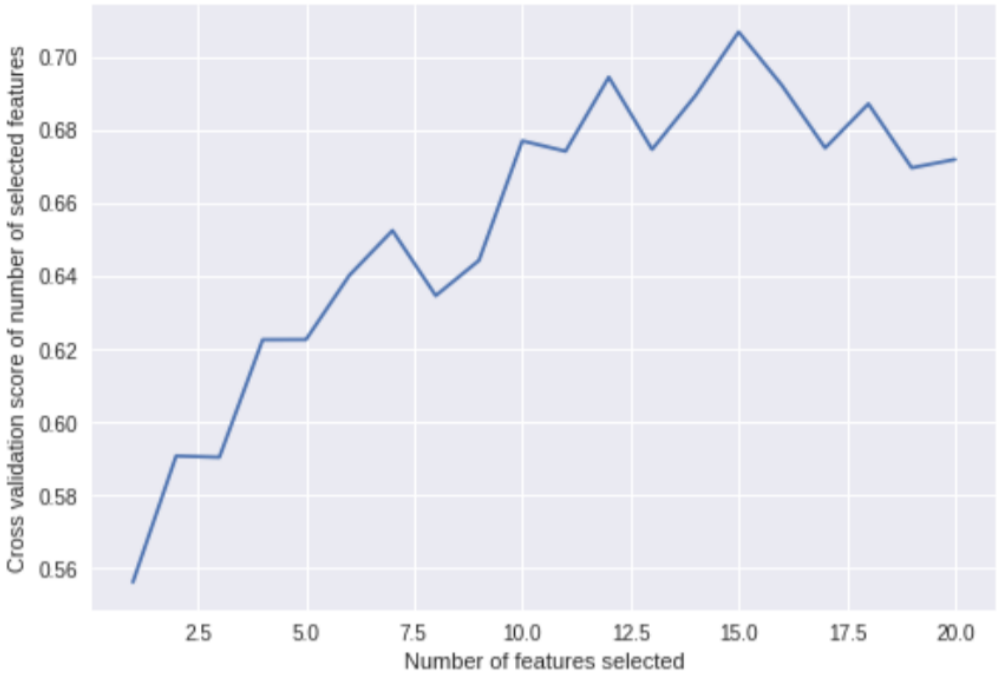


Fig. 4.2 Cross Validation Score For Number of Features Selected

```
RFE cross
Optimal number of features : 15
Best features : Index(['F-1', 'F-2', 'F-3', 'F-4', 'F-6', 'F-9', 'F-10', 'F-11', 'F-12',
       'F-13', 'F-14', 'F-16', 'F-17', 'F-19', 'F-20'],
      dtype='object')
```

Fig. 4.3 Optimal Feature Selection Using RFE Cross Validation

4.5 Accuracy of Different Algorithms for Optimal Features

After finding out the accuracy for the 20 features, we than extracted the optimal features from the features that we used initially. We found out 15 optimal features. We then deleted the respective features that were not need any more. and again ran the 6 algorithms on them.

accuracy:72.43% +/- 6.49%(micro average: 72.44%)			
	true No	true Yes	class precision
pred. NO	286	78	78.57%
pred. Yes	81	132	61.97%
class recall	77.93%	62.86%	

Table 4.7 Accuracy, Precision and Recall Of Deep Learning for Optimal Features

4.5.1 Deep Learning

From Table 4.7 we can see, that our accuracy is 72.43% which approximately 72%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 132 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 213. Therefore dividing 132 by 213 and then multiplying it with 100, we get a precision of 61.97% approximately 62%. Then we look at the recall, which is the number of true positives in actual yes results. There are 132 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 132 by 210 and then multiplying it with 100, we get a recall of 62.86% approximately 63%. Also, We see a False Negative of 78 out of actual 210 depressed cases which is comparatively lower for our system. False Positive is 81 out of actual 367 not depressed cases.Therefore, a good balance in False Positive and False Negative.

4.5.2 Generalized Linear Model

From Table 4.8 we can see, that our accuracy is 74.87% which approximately 75%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 110 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 155. Therefore dividing 110 by 155 and then multiplying it with 100, we get a precision of 70.97% approximately 72%. Then we look at the recall, which is the number of true positives in actual yes results. There are 110 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 110 by 210 and then multiplying it with 100, we get a recall of 52.38% approximately 52%. Also, We see a False Negative of 100 out of actual 210 depressed cases which is comparatively moderate for our system. False Positive is 45 out of actual 367 not depressed cases.

accuracy:74.87% +/- 5.26%(micro average: 74.87%)			
	true No	true Yes	class precision
pred. NO	322	100	76.30%
pred. Yes	45	110	70.97%
class recall	87.74%	52.38%	

Table 4.8 Accuracy, Precision and Recall Of Generalized Linear Model for Optimal Features

accuracy:72.79% +/- 5.01%(micro average: 72.79%)			
	true No	true Yes	class precision
pred. NO	284	74	79.33%
pred. Yes	83	136	62.10%
class recall	77.38%	64.76%	

Table 4.9 Accuracy, Precision and Recall Of Gradient Boosted Algorithm for Optimal Features

4.5.3 Gradient Boosted Algorithm

From Table 4.9 we can see, that our accuracy is 72.79% which approximately 73%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 136 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 219. Therefore dividing 136 by 219 and then multiplying it with 100, we get a precision of 62.10% approximately 62%. Then we look at the recall, which is the number of true positives in actual yes results. There are 136 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 136 by 210 and then multiplying it with 100, we get a recall of 64.76% approximately 65%. We see a False Negative of 74 out of actual 210 depressed cases which is comparatively lower for our system. False Positive is 83 out of actual 367 not depressed cases. Therefore, a good balance in False Positive and False Negative.

4.5.4 K-Nearest Neighbor

From Table 4.10 we can see, that our accuracy is 66.37% which approximately 66%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 67 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 118.

accuracy:66.37% +/- 5.88%(micro average: 66.38%)			
	true No	true Yes	class precision
pred. NO	316	143	68.85%
pred. Yes	51	67	56.78%
class recall	86.10%	31.90%	

Table 4.10 Accuracy, Precision and Recall Of K-Nearest Neighbor for Optimal Features

accuracy:72.77% +/- 3.37%(micro average: 72.79%)			
	true No	true Yes	class precision
pred. NO	307	97	75.99%
pred. Yes	60	113	65.32%
class recall	83.65%	53.81%	

Table 4.11 Accuracy, Precision and Recall Of Random Forest for Optimal Features

Therefore dividing 67 by 118 and then multiplying it with 100, we get a precision of 56.78% approximately 57%. Then we look at the recall, which is the number of true positives in actual yes results. There are 67 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 67 by 210 and then multiplying it with 100, we get a recall of 31.90% approximately 32%. Also, We see a False Negative of 143 out of actual 210 depressed cases which is comparatively a very high for our system. False Positive is 51 out of actual 367 not depressed cases.

4.5.5 Random Forest

From Table 4.11 we can see, that our accuracy is 72.77% which approximately 73%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 113 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 173. Therefore dividing 113 by 173 and then multiplying it with 100, we get a precision of 65.32% approximately 65%. Then we look at the recall, which is the number of true positives in actual yes results. There are 113 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 113 by 210 and then multiplying it with 100, we get a recall of 53.81% approximately 54%. Also, We see a False Negative of 97 out of actual 210 depressed cases which is comparatively moderate for our system. False Positive is 60 out of actual 367 not depressed cases.

accuracy:74.19% +/- 4.44%(micro average: 74.18%)			
	true No	true Yes	class precision
pred. NO	322	104	75.59%
pred. Yes	45	106	70.20%
class recall	87.74%	50.48%	

Table 4.12 Accuracy, Precision and Recall Of Support Vector Machine for Optimal Features

4.5.6 Support Vector Machine

From Table 4.12 we can see, that our accuracy is 74.19% which approximately 74%. Also if we look at the precision, where precision means the total number of true positive in all the prediction of yes. There are 106 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of prediction of yes is 151. Therefore dividing 106 by 151 and then multiplying it with 100, we get a precision of 70.20% approximately 70%. Then we look at the recall, which is the number of true positives in actual yes results. There are 106 individuals who were predicted as yes and they are actually depressed which is the true positive and total number of actual yes is 210. Therefore dividing 106 by 210 and then multiplying it with 100, we get a recall of 50.48% approximately 50%. Also, We see a False Negative of 104 out of actual 210 depressed cases which is comparatively moderate for our system. False Positive is 45 out of actual 367 not depressed cases.

4.6 Comparison between Algorithms for optimal features

We will now compare accuracy, precision, recall and f-measure along with the False Negative and False Positive of all the algorithms applied on the optimal features.. Table 4.4 shows the comparison between all the algorithms using a histogram.

From Figure 4.4 we can see that except K-NN all the algorithms gave a very close accuracy. Here, blue bar represents accuracy, red bar represents precision, recall is represented using green bar and f-measure is represented using purple bar. From the histogram we can see that we have received the highest accuracy, which is 75% in Generalized Linear Model and the lowest accuracy is in K-NN, which is 66%. Even though it seems 66% is not making much of a difference but actually K Nearest Neighbor is not good for our system and this clearly indicated by the f-measure value, which is only 40%. F-measure is more accurate and it measure's the test's accuracy and is calculated using weighted harmonic mean of precision

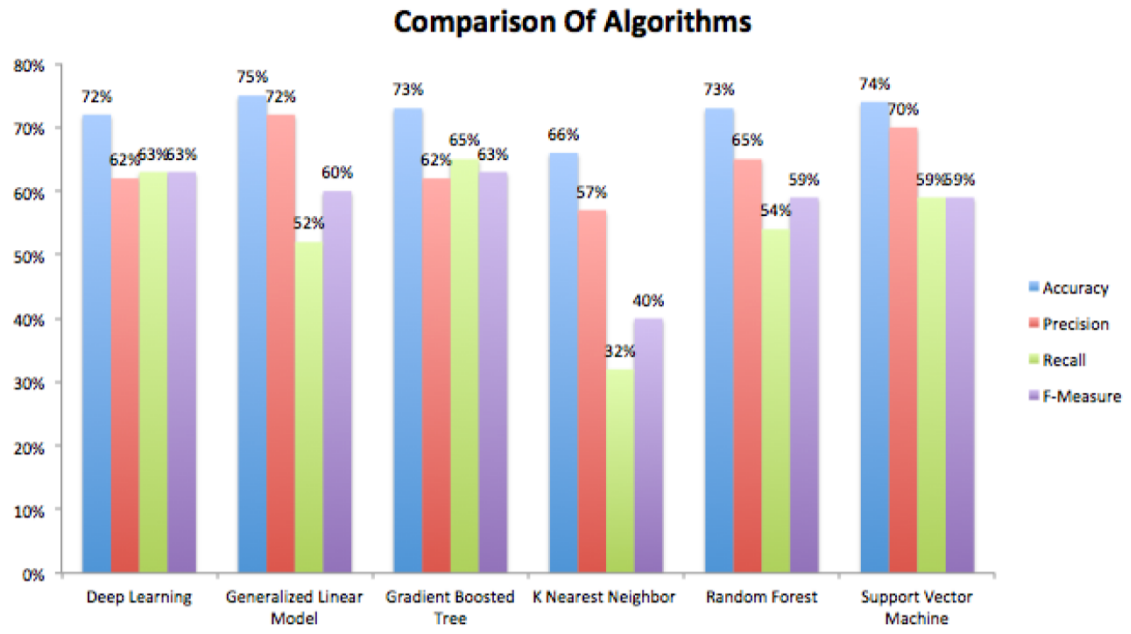


Fig. 4.4 Comparison Between Algorithms For Optimal Features

and recall. The higher the f-measure, the better our system will be. Thus K-NN is a bad choice due to less f-measure value. The reason behind this is because, K-NN works best with fewer features when classifying. However, when we trained our model, we converted this 15 features using dummy variable finally giving 38 features, which is too much for a K-NN algorithm because in higher dimensions the points are too close to each other and hence difficult to assume from the graph. Also, there is a huge difference in the precision and recall. Precision is still 57%, which is not bad enough whereas recall is only 32%, which shows that only 32% of the actual depressed case has been correctly identified.

All the other algorithms are also giving good results. SVM has accuracy of 74% with a f-measure of 59%. Generalized Linear Model is giving an accuracy of 75% and a f-measure value of 60%. Deep Learning is giving an accuracy of 72% and a f-measure value of 63%. Gradient Boosted Tree is giving a accuracy of 73% and f-measure of 63%. From the given accuracy and f-measure we can see all of them are more or less close to each other except K-NN. Therefore, K-NN is not at all a good idea for our system.

Moreover, we can see that there is a good balance in False Positive and False Negative and these values are comparatively lower for the algorithms in Deep Learning and Gradient Boosted Tree. Also, from the prediction, accuracy and f-measure, we can come to a con-

clusion that Deep Learning and Gradient Boosted Tree gives the best result for our model for both the cases we have seen. This can be clearly seen as both of them have the highest f-measure of 63%, which is the weighted average of precision and recall. We can see that precision for both of this algorithm is 62% and recall is 63% for Deep Learning and 65% for Gradient Boosted Tree. There is very less difference in precision and recall for both of this algorithm whereas for other algorithms there is a good difference in precision and recall. 72% precision and 52% recall for Generalized Linear Model. 65% precision and 54% recall for Random Forest. 70% precision and 59% recall for Support Vector Machine. Therefore, we can conclude that Deep Learning and Gradient Boosted Tree is better compared to others for our system when we use the optimal features.

We will now see a comparison between the algorithms used for the 20 features and the selected optimal features in the next section.

4.7 20 Features Versus Optimal 15 Features Comparison

In Table 4.5, Table 4.6, Table 4.7 and Table 4.8 the blue bar represents prediction result for 20 features and red bar represents prediction result for the optimal 15 features.

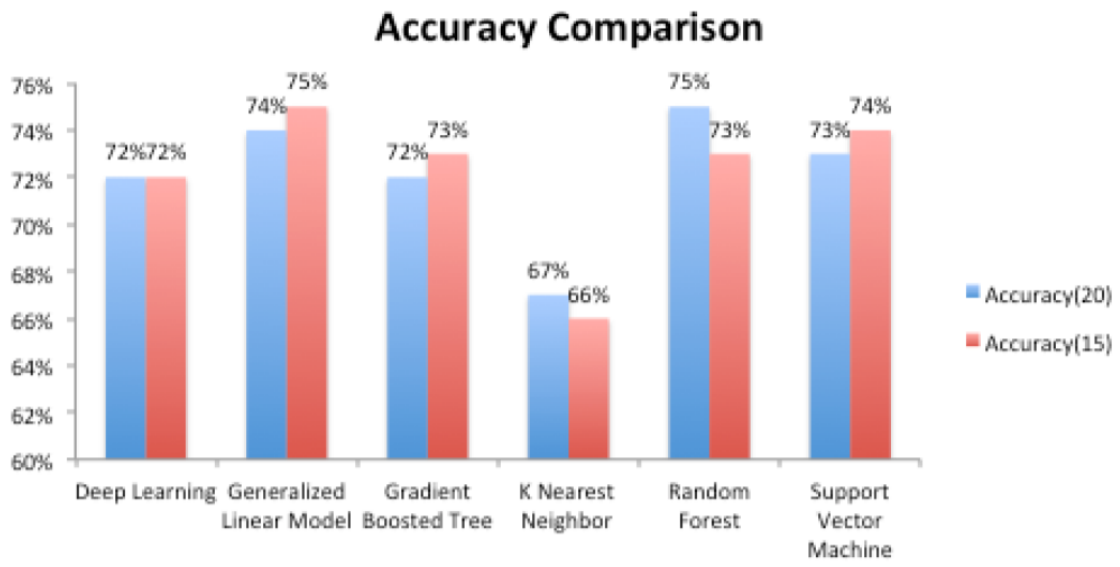


Fig. 4.5 Accuracy Comparison Between 20 Features And Optimal Features

From Figure 4.5 we can see from the accuracy comparison that there was no change in

the overall accuracy in Deep Learning. The accuracy remained at 72%. However, the accuracy increased from 74% to 75% in generalized linear model for the optimal features. The accuracy for Gradient Boosted Tree increased from 72% to 73% for the optimal features and for support vector machine it increased from 73% to 74%. Only for Random forest it decreased to 73% from 75% when we used the optimal features. Also K-NN had a significant drop of 1% that is from 67% to 66%.

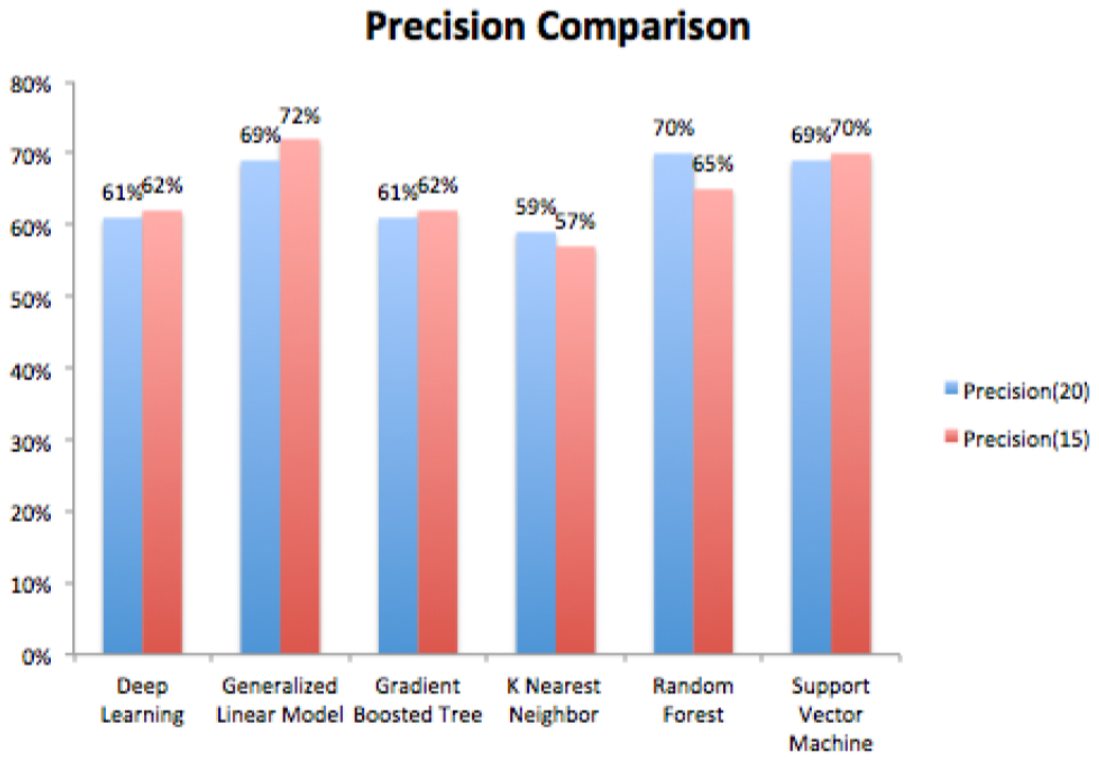


Fig. 4.6 Precision Comparison Between 20 Features And Optimal Features

From Figure 4.6 we can see from the precision comparison that there was a change in the overall precision in Deep Learning. It increased from 61 to 62% thus it shows that deep learning is helping us get a very good result for our system. The precision increased from 69% to 72% in generalized linear model for the optimal features. Thus having a significant rise in the precision level. The precision for Gradient Boosted Tree increased from 61% to 62% for the optimal features and for support vector machine it increased from 69% to 70%. Only for Random forest it decreased to 65% from 70% when we used the optimal features, which is a big drop in the precision level. Also K-NN had a significant drop of 2% that is from 59% to 57%.

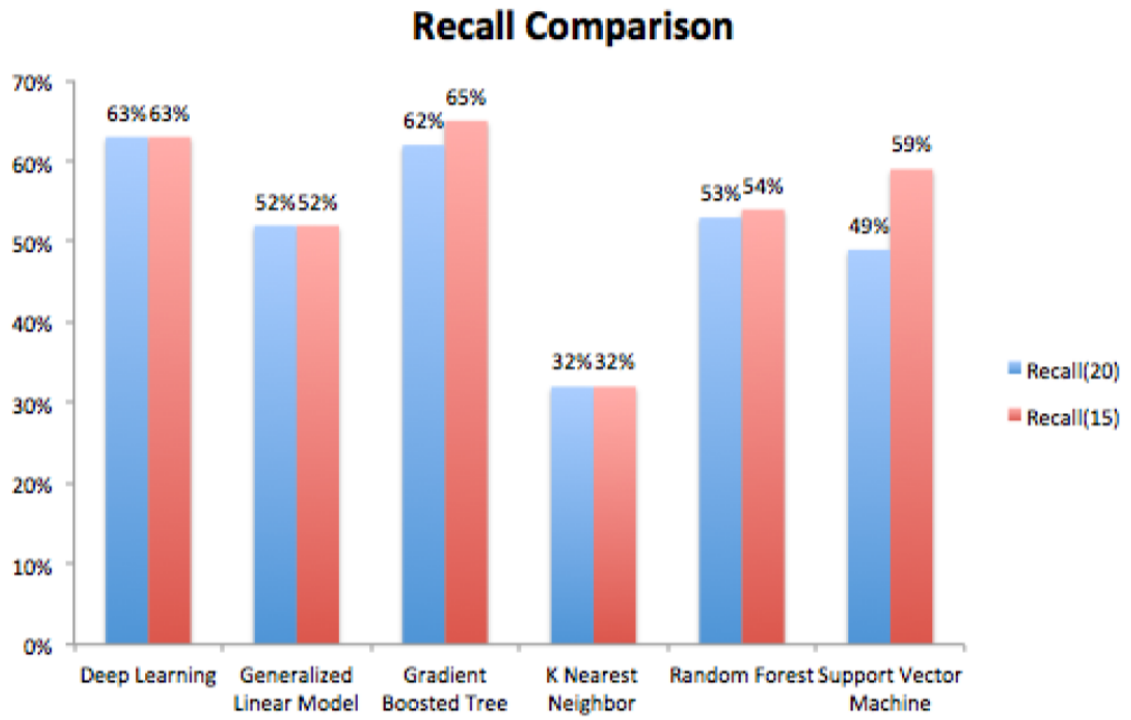


Fig. 4.7 Recall Comparison Between 20 Features And Optimal Features

From Figure 4.7 we can see from the recall comparison that there was no change in the overall recall in Deep Learning. The recall remained at 63%. The recall was also same for Generalized linear model. It was 52%. However, the recall increased from 62% to 65% in Gradient Boosted when optimal features were used and for support vector machine it increased from 49% to 59%. This rise in recall shows how important it was to use the 15 features as it will help reduce over fitting of data. Also for Random forest it increased to 54% from 53% when we used the optimal features. Also there was no change for K-NN algorithm. It was 32% for both the case.

From Figure 4.8 we can see from the F-measure comparison that there was a change in the overall f-measure in Deep Learning. It increased from 62 to 63% thus it shows that deep learning is helping us get a very good result for our system. The f-measure increased from 59% to 60% in generalized linear model for the optimal features. The f-measure for Gradient Boosted Tree increased from 62% to 63% for the optimal features and for support vector machine it increased from 58% to 59%. Only for Random forest it decreased to 59% from 60% when we used the optimal features. Also K-NN had a significant drop of 1% that is from 41% to 40%.

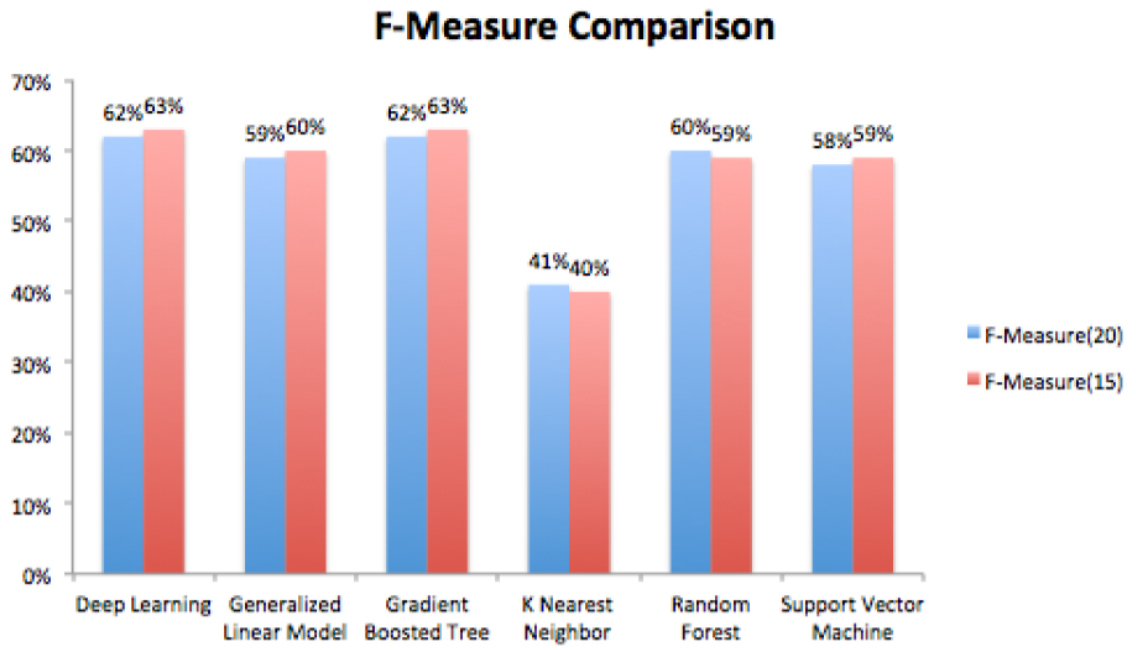


Fig. 4.8 F-Measure Comparison Between 20 Features And Optimal Features

Therefore to conclude, after looking at the comparison of accuracy, precision, recall and the f-measure value we can come to a conclusion that using the optimal 15 features is feasible for our model. This not only reduces over fitting of data but also reduces the dimension of the data. Also we can say that our prediction does not depend on the answer of the questions that were removed after UFE with cross validation and random forest. For 4 of the algorithm, which are deep learning, generalized machine learning, gradient boosted tree and support vector machine we saw how the data value changed when we used the optimal features. They showed even a better performance as we can see from the accuracy comparison. Therefore, since for both the cases we can see that for deep learning and gradient boosted tree we get the highest f-measure, we will be using this 2 algorithms in the near future and research more on them and further extend our work so that we can increase our accuracy so that we can help the young generation to lead a beautiful stress free life.

Chapter 5

Final Remarks

This chapter concludes by restating the problem, the work done in this research and improvements and further work to be done in this field.

5.1 Conclusion

The most prevalent mental disorder or illness, depression is a broad area of research that has a lot of implications in medical and psychology. Young people are more likely to suffer from this disease especially residents of lower middle income countries due to various socio-demographic reasons. Depression has its roots deeply ingrained in the lifestyle, habits and behavior of people. This research sought to exploit this relation by using some social and personal data in order to predict depression in individuals. There has not been much work on analyzing depression among Bangladesh citizens. However, this is one of the major social and medical problems that is continuously growing and hence need to be addressed immediately. There has been a recent rise in suicide rates across the country, majority of which was a result of depression as stated in newspaper articles. Moreover, most of these reported suicide cases were university students including one from BRAC University and several from Dhaka University, two of the top educational institutions in the country. One of the main objectives of this study was to identify early cases of depression and ask individuals at risk of developing depression to consult a psychiatrist or consultant. Also, professionals can use our system to identify causes of depression in specific individuals. Random Forest Algorithm was used to find out relevant features that contribute to depression and the six algorithms were used to predict depression in undergraduate university students of Bangladesh based on the selected features. Deep Learning and Gradient Boost Algorithm proved to be best methods for predicting depression. The need for identifying depression early is dire as the disorder can worsen very quickly. The proposed model could be used by

psychologists, counsellors, universities to find out depression in students so that appropriate steps can be taken to reduce the effect and impact of the disease and help young people live a healthier life, a happier life.

5.2 Limitations and Future work

Limitations:

- From the experience from the data collection process and the feedback from the participant we have find out that participant was a little bit reluctant to fill this long survey paper which make it difficult for the participant to hold their concentration fully throughout the whole survey process.
- Number of non depressed and moderately depressed people outnumbered the severe and extreme depression people in our research. So there are less severe depressed data set to train our model compared to non depressed or moderate level depression data set.
- The number of relevant reasons are so selective and was unable to add less important relevant features in our survey form because that may result our research paper to be more lengthy which may hamper the credibility of the data collection process.
- We were only able to cover one university where we were able to take our survey with one to one mentoring.

In future we want to make our model more accurate and find out more relevant reasons or features which are related to depression in university going students. Currently we are continuing to collect more data from online source. Beside this we made our survey form more precise by removing the bangla depression scale which will reduce the completion time for this survey. Along with this we have selected more relevant reasons which is related to depression so that there is an improvement in data collection process,accuracy of the model and can find more relevant and irrelevant reasons.

Beside this we are planning to cover more universities of Bangladesh both public and private universities where we will collect data in face to face survey taking process. To get more extreme,and severe depression datasets we are combinedly working together with brac u counselor so that we get more extreme and severe depressed data instances .

We are planning to make a system for BRAC University counselor which will be able to predict the depression in university students by using our system and it will be able to help the counsellor to detect the relevant reasons for which university going students are becoming victims of depression. Students can also furthermore self evaluate themselves whether they are depressed or not and seek help if necessary. Finally, for future we are planning to make a system, so that answers to the features of our system can be easily extracted from different social networking sites and other places for every students so that they can be kept under regular observation. This will help universities to understand whether a student requires immediate attention or not and counsel accordingly to save him/her from depression as it is a very serious condition if overlooked.

References

[Rel] Relationships and depression | relate. [note] https://www.relate.org.uk/relationship-help/help-relationships/mental-health/relationships-and-depression.

[2] (1996). Beck depression inventory®-ii. [online] https://www.pearsonclinical.com/psychology/products/100000159/beck-depression-inventoryii-bdi-ii.html#tab-details.

[3] (2018). Nimh » depression. [online] https://www.nimh.nih.gov/health/topics/depression/index.shtml.

[4] Ahmad, N., Hussain, S., and Munir, N. (2018). Social networking and depression among university students. *Pakistan Journal of Medical Research*, 57(2).

[5] Akbayrak, S. (2018). Generalized linear models – towards data science. [online] https://towardsdatascience.com/generalized-linear-models-8738ae0fb97d.

[6] Alexander, R. and Krans, B. (2016). Anxiety, depression & suicide: the lasting effects of bullying. [online] https://www.healthline.com/health-news/bullying-affects-victims-and-bullies-into-adulthood-022013.

[7] Baddele, J. (2009). How to help a depressed friend (and when to stop trying): part 2 | psychology today. [online] https://www.psychologytoday.com/us/blog/embracing-the-dark-side/200906/how-help-depressed-friend-and-when-stop-trying-part-2.

[8] Bhakta, I. and Sau, A. (2016). Prediction of depression among senior citizens using machine learning classifiers. *International Journal of Computer Applications*, 144(7):11–16.

[9] BHOPB (2016). Are depression and social media usage linked? - behavioral health of the palm beaches. [online] https://www.bhpalmbeach.com/are-depression-and-social-media-usage-linked/.

[10] Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., and Hanson, C. L. (2016). Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR mental health*, 3(2).

[11] Bronshtein, A. (2017). A quick introduction to k-nearest neighbors algorithm. [online] https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7.

[12] Brownlee, J. (2016a). A gentle introduction to the gradient boosting algorithm for machine learning. [online] https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/.

[13] Brownlee, J. (2016b). What is deep learning? [online] https://machinelearningmastery.com/what-is-deep-learning/.

[Butler] Butler, A. Interpersonal & intrapersonal conflict | livestrong.com. [note] https://www.livestrong.com/article/144539-interpersonal-intrapersonal-conflict/?ajax=1&amp%3Bis=1, month = , year = ,.

[15] Campbell, J. (2016). Financial stress leads to symptoms of depression, ptsd. [online] https://www.moneymanagement.org/blog/2016/05/financial-stress-leads-to-symptoms-of-depression.

[16] Cristianini, N. and Shawe-Taylor, J. (2014). *An Introduction to support vector machines: and other kernel-based learning methods*. Cambridge University Press.

[17] Daimi, K. and Banitaan, S. (2014). Using data mining to predict possible future depression cases. *International Journal of Public Health Science (IJPHS)*, 3(4):231–240.

[18] DATAI (2018). Feature selection and data visualization | kaggle. [online] https://www.kaggle.com/kanncaa1/feature-selection-and-data-visualization?fbclid=IwAR1cRZO2FWLqSRq0NqLmSchYsihTtvNF29PqbkH3bpi-yiIPobfMn3Wg8Q.

[19] Deykin, E. Y., Levy, J. C., and Wells, V. (1987). Adolescent depression, alcohol and drug abuse. *American Journal of Public Health*, 77(2):178–182.

[20] Donges, N. (2018). The random forest algorithm – towards data science. [online] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd.

[DualDiagnosis.org] DualDiagnosis.org. Depression and addiction, dual diagnosis. [online] https://www.dualdiagnosis.org/depression-and-addiction/.

[22] Fekkes, M., Pijpers, F. I., and Verloove-Vanhorick, S. P. (2004). Bullying behavior and associations with psychosomatic complaints and depression in victims. *The Journal of pediatrics*, 144(1):17–22.

[23] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[24] Gandhi, R. (2018). Support vector machine — introduction to machine learning algorithms. [online] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[25] Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

[26] Grover, P. (2017). Gradient boosting from scratch – ml review – medium. [online] https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.

[27] Hafen, C. A., Allen, J. P., Schad, M. M., and Hessel, E. T. (2015). Conflict with friends, relationship blindness, and the pathway to adult disagreeableness. *Personality and individual differences*, 81:7–12.

[28] Han, J., Rodriguez, J. C., and Beheshti, M. (2008). Diabetes data analysis and prediction model discovery using rapidminer. In *2008 Second International Conference on Future Generation Communication and Networking*, pages 96–99. IEEE.

[29] Hiremath, R. C. and Debaje, S. P. (2017). Assessment of prevalence of domestic violence and mental health profile of adolescents exposed to domestic violence in an urban slum in mumbai. *International Journal of Research in Medical Sciences*, 2(1):290–292.

[30] Houle, J. N., Staff, J., Mortimer, J. T., Uggen, C., and Blackstone, A. (2011). The impact of sexual harassment on depressive symptoms during the early occupational career. *Society and mental health*, 1(2):89–105.

[31] Hysenbegasi, A., Hass, S. L., and Rowland, C. R. (2005). The impact of depression on the academic productivity of university students. *Journal of Mental Health Policy and Economics*, 8(3):145.

[32] Ibrahim, A. K., Kelly, S. J., Adams, C. E., and Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. *Journal of psychiatric research*, 47(3):391–400.

[33] Iliades, C. (2018). 9 different types of depression | everyday health. [online] https://www.everydayhealth.com/depression-pictures/different-types-of-depression.aspx.

[34] Jain, S. (2017). A comprehensive beginners guide for linear, ridge and lasso regression. [online] https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/.

[35] January, J., Madhombiro, M., Chipamaunga, S., Ray, S., Chingono, A., and Abas, M. (2018). Prevalence of depression and anxiety among undergraduate university students in low-and middle-income countries: a systematic review protocol. *Systematic reviews*, 7(1):57.

[36] Kaltiala-Heino, R., Rimpelä, M., Marttunen, M., Rimpelä, A., and Rantanen, P. (1999). Bullying, depression, and suicidal ideation in finnish adolescents: school survey. *Bmj*, 319(7206):348–351.

[37] Kaplow, J. B., Saunders, J., Angold, A., and Costello, E. J. (2010). Psychiatric symptoms in bereaved versus nonbereaved youth and young adults: a longitudinal epidemiological study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(11):1145–1154.

[38] Karunakaran, D. (2018). Deep learning series 1: Intro to deep learning – intro to artificial intelligence – medium. [online] https://medium.com/intro-to-artificial-intelligence/deep-learning-series-1-intro-to-deep-learning-abb1780ee20.

[39] Keller, J. M., Gray, M. R., and Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (4):580–585.

[40] Keyes, K. M., Pratt, C., Galea, S., McLaughlin, K. A., Koenen, K. C., and Shear, M. K. (2014). The burden of loss: unexpected death of a loved one and psychiatric disorders across the life course in a national study. *American Journal of Psychiatry*, 171(8):864–871.

[41] Khadr, S., Clarke, V., Wellings, K., Villalta, L., Goddard, A., Welch, J., Bewley, S., Kramer, T., and Viner, R. (2018). Mental and sexual health outcomes following sexual assault in adolescents: a prospective cohort study. *The Lancet Child & Adolescent Health*, 2(9):654–665.

[42] Khandelwal, R. (2018). K fold and other cross-validation techniques – data driven investor – medium. [online] https://medium.com/datadriveninvestor/k-fold-and-other-cross-validation-techniques-6c03a2563f1e.

[43] Khurshid, S., Parveen, Q., Yousuf, M. I., and Chaudhry, A. G. (2015). Effects of depression on students' academic performance. *Science International*, 27(2):1619–1624.

[44] Koehrsen, W. (2017). Random forest simple explanation – william koehrsen – medium. [online] https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d.

[45] Kumar, G. S., Jain, A., and Hegde, S. (2012). Prevalence of depression and its associated factors using beck depression inventory among students of a medical college in karnataka. *Indian journal of Psychiatry*, 54(3):223.

[46] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

[47] Lison, P. (2015). "an introduction to machine learning.

[48] Merz, B. (2017). Six common depression types - harvard health. [online] urlhttps://www.health.harvard.edu/mind-and-mood/six-common-depression-types.

[49] Nall, R. (2017). Domestic violence and abuse - the impact on children and adolescents | royal college of psychiatrists. [online] https://www.rcpsych.ac.uk/mental-health/parents-and-young-people/information-for-parents-and-carers/domestic-violence-and-abuse-effects-on-children.

[50] Nambisan, P., Luo, Z., Kapoor, A., Patrick, T. B., and Cisler, R. A. (2015). Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 2906–2913. IEEE.

[51] Nelder, J. A. and Baker, R. J. (2004). Generalized linear models. *Encyclopedia of statistical sciences*, 4.

[52] Ngo, S. (2016). 5 signs that you are under too much financial stress. [note] https://www.cheatsheet.com/money-career/signs-that-you-are-under-too-much-financial-stress.html/.

[53] Okun, B. and Nowinski, J. (2015). Can grief morph into depression? - harvard health blog - harvard health publishing. [online] https://www.health.harvard.edu/blog/can-grief-morph-into-depression-201203214511.

[54] Oldt, A. (2016). Family history of depression doubles risk for depression. [online] https://www.healio.com/psychiatry/depression/news/online/%7B133607cf-45bc-40a2-89af-80a80a4d29e3%7D/family-history-of-depression-doubles-risk-for-depression.

[55] Parekh, R. (2017). What is depression? [online] https://www.psychiatry.org/patients-families/depression/what-is-depression.

[56] Passos, I. C., Mwangi, B., Cao, B., Hamilton, J. E., Wu, M.-J., Zhang, X. Y., Zunta-Soares, G. B., Quevedo, J., Kauer-Sant'Anna, M., Kapczinski, F., et al. (2016). Identifying a clinical signature of suicidality among patients with mood disorders: a pilot study using a machine learning approach. *Journal of affective disorders*, 193:109–116.

[57] Pietrangelo, A. (2018). Types of depression: 9 forms of depression and their symptoms. [online] https://www.healthline.com/health/types-of-depression.

[58] Post, R. M., Altshuler, L. L., Kupka, R., McElroy, S. L., Frye, M. A., Rowe, M., Grunze, H., Suppes, T., Keck Jr, P. E., Leverich, G. S., et al. (2018). Multigenerational transmission of liability to psychiatric illness in offspring of parents with bipolar disorder. *Bipolar disorders*.

[59] Powers, A. (2018). Ai can identify depression based on a natural conversation, an mit study finds. [online] https://www.forbes.com/sites/annapowers/2018/09/30/ai-senses-depression-in-people-based-on-how-they-talk-an-mit-study-finds/#7fa2763876f5.

[60] Ray, S. (2017). Understanding support vector machine algorithm from examples (along with code). [note] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

[61] Reilly, K. (2018). Anxiety and depression: More college students seeking help | time. [online] http://time.com/5190291/anxiety-depression-college-university-students/.

[62] Report, F. O. (2018). Social media users 30 million in bangladesh: Report. [online] https://thefinancialexpress.com.bd/sci-tech/social-media-users-30-million-in-bangladesh-report-1521797895.

[63] Richardson, T., Elliott, P., Roberts, R., and Jansen, M. (2017). A longitudinal study of financial difficulties and mental health in a national sample of british undergraduate students. *Community mental health journal*, 53(3):344–352.

[64] Rosenberg, D. (2018). 1 in 5 college students have anxiety or depression. here's why. [online] http://theconversation.com/1-in-5-college-students-have-anxiety-or-depression-heres-why-90440.

[65] Sack, D. (2014). Tough truths you should know about addiction, depression | psychology today. [online] https://www.psychologytoday.com/us/blog/where-science-meets-the-steps/201408/tough-truths-you-should-know-about-addiction-depression.

[66] Sammy (2018). K-fold cross validation – sammy – medium. [online] https://medium.com/@shivam.somani09/k-fold-cross-validation-cb85632dba3.

[67] Schimelpfening, N. (2018). 7 most common types of depression. [online] https://www.verywellmind.com/common-types-of-depression-1067313.

[68] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[69] Sharma, P. (2018). Introduction to neural networks, deep learning (deeplearning.ai course). [online] https://www.analyticsvidhya.com/blog/2018/10/introduction-neural-networks-deep-learning/.

[70] Slopen, N., Fitzmaurice, G. M., Williams, D. R., and Gilman, S. E. (2012). Common patterns of violence experiences and depression and anxiety among adolescents. *Social psychiatry and psychiatric epidemiology*, 47(10):1591–1605.

[71] Smith, K. (2018). Substance abuse and depression: A dangerous downward-spiral. [online] https://www.psycom.net/depression-substance-abuse.

[72] Soni, D. (2018). Introduction to k-nearest-neighbors – towards data science. [online] https://towardsdatascience.com/introduction-to-k-nearest-neighbors-3b534bb11d26.

[73] Spector, N. (2017). The hidden health effects of sexual harassment. [online] https://www.nbcnews.com/better/health/hidden-health-effects-sexual-harassment-ncna810416.

[74] Srivastava, T. (2018). Introduction to knn, k-nearest neighbors : Simplified. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.

[75] Tugend, A. (2017). A climb out of depression, doubt and academic failure - the new york times. [online] https://www.nytimes.com/2017/06/07/education/one-student-tells-her-story-of-a-climb-out-of-depression-.html.

[76] Volkow, N. D. (2004). The reality of comorbidity: depression and drug abuse. *Biological psychiatry*.

[77] Weissman, M. M., Berry, O. O., Warner, V., Gameroff, M. J., Skipper, J., Talati, A., Pilowsky, D. J., and Wickramaratne, P. (2016). A 30-year study of 3 generations at high risk and low risk for depression. *JAMA psychiatry*, 73(9):970–977.

[78] Weissman, M. M., Wickramaratne, P., Adams, P., Wolk, S., Verdeli, H., and Olfson, M. (2000). Brief screening for family psychiatric history: the family history screen. *Archives of General Psychiatry*, 57(7):675–682.

[79] Whitbourne, S. K. (2016). How being depressed can affect your relationships | psychology today. [online] https://www.psychologytoday.com/us/blog/fulfillment-any-age/201602/how-being-depressed-can-affect-your-relationships.

[80] Wolke, D. and Lereya, S. T. (2015). Long-term effects of bullying. *Archives of disease in childhood*, 100(9):879–885.

[81] Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846.

[82] Zeltzer, L. (2008). Beck depression inventory (bdi, bdi-ii) - stroke engine. [online] https://www.strokengine.ca/en/assess/bdi/.