DETECTION OF DEPRESSION AND ANXIETY SYMPTOMS IN UNIVERSITY STUDENTS USING MACHINE LEARNING

BY

Tahmina Akter
ID: 0242310005103018

This Report Presented in Partial Fulfillment of the Requirements for

The Degree of Masters of Science in Computer Science and Engineering

Supervised By

Professor Dr. Md. Ismail Jabiullah

Professor

Department of Computer Science and Engineering

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

6 JULY 2024

APPROVAL

This Project/Thesis titled **"Detection Of Depression And Anxiety Symptoms In University Students Using Machine Learning"**, submitted by **Tahmina Akter,** ID No: **0242310005103018** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 06-07-2024.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori, PhD — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zahid Hasan, PhD — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Arif Mahmud, PhD — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zulfiker Mahmud — **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Professor Dr. Md. Ismail Jabiullah, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Professor Dr. Md. Ismail Jabiullah
Professor
Department of CSE
Daffodil International University

Submitted by:

Tahmina

Tahmina Akter
ID: 0242310005103018
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We really grateful and wish our profound our indebtedness to **Professor Dr. Md. Ismail Jabiullah**, Professor, Department of CSE Daffodil International University, Dhaka. His endless patience , scholarly guidance , continual encouragement , constant and energetic supervision, constructive criticism , valuable advice , reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, PhD Professor and Head, Department of CSE**, for his kind help to finish our thesis and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and passion of our parents.

ABSTRACT

This report explores the application of machine learning techniques for the detection of depression and anxiety symptoms among university students. The prevalence of mental health issues in this demographic poses significant challenges to academic success and overall well-being. Traditional methods of identifying and addressing these concerns often fall short due to limitations such as stigma, subjective assessments, and resource constraints. Using diverse datasets encompassing demographic information, academic performance metrics, social media activity, and smartphone usage patterns, machine learning models are trained to recognize patterns indicative of mental health issues. The study evaluates various machine learning algorithms, including Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), k-Nearest Neighbors (KNN), and Naive Bayes (NB), to determine their effectiveness in predicting depression and anxiety symptoms. Ethical considerations such as data privacy, bias mitigation, and responsible model deployment are also addressed. The findings offer insights into the potential of machine learning to revolutionize mental health assessment in university settings, providing opportunities for early intervention and personalized support for students.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

In past few years, mental health issues among university students have garnered increasing attention worldwide. The change to life, study pressures, social responsibilities, and expectations often contribute to heightened levels of stress, anxiety, and depression among this age group. Depression is an increasingly significant concern that profoundly impacts an individual's lifestyle, disrupting normal functioning and hindering perspectives [1]. Identifying and addressing these mental health concerns early on is crucial for promoting well-being and academic success.

Traditional methods of detecting depression and anxiety symptoms rely heavily on self-reporting or clinical assessments, which may be subjective, time-consuming, and resource-intensive. However, advancements in technology, especially in machine learning, offer promising avenues for more efficient and accurate detection methods.

This report discusses the use of algorithms to train models in the detection of depression symptoms among university students. By leveraging data collected from various sources such as student health surveys, electronic medical records, social media activity, and smartphone usage patterns, machine learning models are trained to detect trends indicative of psychological concerns.

The main reason for the study is the creation of a model which is capable of identifying students at risk of depression based on their digital footprint and behavioral patterns. By analyzing diverse datasets encompassing demographic information, academic performance metrics, social interactions, and psychological assessments, we aim to create a good framework for early detection and intervention.

Additionally, this report examines the ethical considerations surrounding the use of machine learning in mental health assessment, including privacy of information, bias reduction, and accountable deployment of predictive models. Ensuring transparency, accountability, and user consent are paramount in the development and implementation of such technologies.

By harnessing the power of algorithms, universities and mental health counselors can enhance their ability to support students' mental well-being proactively. Through timely identification, personalized interventions, and ongoing monitoring, we can strive towards

creating a campus environment that fosters resilience, empathy, and holistic health among its student population.

1.2 Motivation

The motivation for this analysis comes from the alarming rise in mental health issues among university students worldwide, particularly concerning depression. Research has shown that the transition to university life, coupled with academic pressures and social challenges, often exacerbates these conditions, leading to adverse effects on students' well-being and academic performance. Traditional methods of addressing mental health concerns in university settings have proven inadequate, with barriers such as stigma, limited resources, and subjective assessments hindering effective intervention. Thus, there is a critical need for innovative approaches that can proactively identify and support students experiencing depression and anxiety, ultimately creating a better and healthier learning place.

The advancement in machine learning technology presents a promising opportunity to revolutionize mental health care in university settings. By harnessing the power of Analysis of data, statistical modeling, and training algorithms can analyze diverse datasets encompassing students' behavioral patterns, academic performance metrics, and digital footprints to identify early indicators of depression and anxiety. This data-driven approach offers several advantages, including increased objectivity, scalability, and accessibility compared to traditional methods. By developing machine learning-based tools for detecting mental health symptoms, we aim to empower universities and mental health practitioners with the means to provide timely interventions and personalized support, ultimately improving the overall well-being and academic success of those who study in university.

1.3 Problem Definition

The problem elaborated in this report regards to improving the detection of depression and anxiety symptoms among university students using machine learning techniques. Traditional methods of identifying mental health issues in this demographic often rely on subjective self-reporting or resource-intensive clinical assessments, leading to delay in diagnose and delayed intervention. By using the capabilities of machine learning, we aim to develop more efficient, objective, and scalable approaches to identify at-risk students early on which will allow for timely intervention and support.

1.4 Research Questions

The thesis questions which we discussed for our study are mentioned in the below:

1. How can machine learning algorithms effectively identify depression and anxiety symptoms in university students?
2. Which combination of data can give the most accurate predictions of mental health issues among university students?
3. What are the main challenges and limitations associated with using machine learning for mental health detection in university students?
4. How do different machine learning algorithms compare in terms of their performance for detecting depression and anxiety symptoms?
5. How can we ethically implement machine learning-based detection for mental health assessment in universities?

1.5 Research Methodology

The way we want will perform the study is called research methodology. The methodology covers multiple steps which are discussed below.

Data Collection:

We will gather information from student surveys, social media, smartphone usage, and academic records, ensuring participants' privacy and consent.

Data Cleaning and Preprocessing:

We'll clean the data to remove errors, handle missing values, and standardize formats. We'll also explore the data to understand its characteristics.

Feature Selection:

We'll pick the most relevant features for our models using methods such as correlational evaluation and emphasis on features ranking.

Model Training and Evaluation:

We will use the cleaned and preprocessed data to teach various training machines models, including Support Vector Machines (SVM), Decision Trees (DT),Random Forests (RF),k-Nearest Neighbors (KNN),and Naive Bayes. All of the strategies will be trained with part of

the data and assessed using relevant performance measures include proficiency,precision, and recall, and matrix. Many techniques will be used to ensure that the models are accurate and good.

Model Selection and Optimization:

After performance metrics are obtained during evaluation, we will select the most suitable machine learning model(s) for detecting depression and anxiety symptoms in university students. We will further optimize the selected model(s) by tuning hyper parameters, adjusting feature representations, and exploring ensemble methods to enhance predictive accuracy and reliability.

Ethical Considerations:

We'll ensure all procedures follow ethical guidelines, respecting participant privacy and addressing biases and limitations transparently.

1.6 Report Layout

Chapter 1, we discuss the introduction, motivation, problem description, questions related to research, and methodology for the research. Introduction will introduce the idea, motivation will show what motivates us, problem definition will be used to define the problem, research questions are used to specify the questions which are related and finally methodology for the study will be discussed here.

Chapter 2, the history of machine learning methods for identifying signs of anxiety and depression, as well as relevant literature, goals, and scope and limitations, will all be covered here. We will discuss deferent related works and how they relate to our work. We will also discuss goals for the study and what are the scopes of the study. The limitations will be shared here also.

Chapter 3, we covered the method or workflow is thoroughly in chapter 3.The methodology of the investigation will be described in this part. Our discussion will include how we gathered and cleaned the data. The preprocessing methods will also be discussed here.

Chapter 4, all the findings of the research will be presented by us here. We have presented the findings from every classifier we have employed in this study in this chapter. We will discuss the classifier use in context of our study and how they are useful for our study.

Chapter 5, here we will discuss how this research might be applied in universities. The significance of this task and its ethical implementation will also be discussed here.

Chapter 6, the report's conclusion is included here. The performance summary of the model is given in this section. An accuracy comparison is also given in this section. The implementation and outputs of the models are also discussed in this section. The chapter finishes in summary of how the work is limited. Future work information was also discussed in this chapter.

Chapters 7, all of the references that we included in our research are listed in this chapter which follows IEEE rules.

CHAPTER 2

BACKGROUND

2.1 Introduction

Similar to a lot of other places on this planet; mental health and psychological problems among university students have increasingly become a matter of concern in Bangladesh. The transition from school to university life, coupled with academic pressures, societal expectations, and personal challenges, often exacerbates conditions such as depression and anxiety among students [2]. However, identifying and addressing these mental health issues in a timely manner remains a significant challenge due to various factors, including stigma, limited access to mental health resources, and cultural barriers to seeking help.

The ability of machine learning to help in regards to the detection of depression symptoms within those who study in Bangladeshi universities is very relevant. Machine learning algorithms offer a data-driven approach to mental health assessment, capable of analyzing diverse datasets comprising students' demographic information, academic performance metrics, social interactions, and digital footprints [3]. By leveraging this wealth of data, machine learning models have the potential to identify early indicators of mental health issues, thereby enabling timely intervention and support for students in need.

However, there are some limits to using machine learning for psychological assessments in Bangladesh. Cultural nuances, language barriers, and access to technology may influence the effectiveness and accessibility of machine learning-based approaches in this context [4]. Additionally, ethical considerations such as privacy for personal data, bias reduction, and the accountable deployment of predictive models must be carefully addressed to ensure the ethical use of these technologies in supporting the mental well-being of university students in Bangladesh. Here, we look at the ways machine learning can help solve these issues and contribute to more effective, inclusive, and culturally appropriate ways for the identification and management of depression symptoms among students in Bangladeshi Universities.

2.2 Related Works

Lots of studies have explored the application of training models with algorithms techniques for detecting depression and anxiety symptoms among university students, providing valuable insights into the feasibility and effectiveness of these approaches.

Minji Gil, Suk-Sun Kim, and Eun Jeong Min conducted a study in 2022 that used machine learning algorithms to predict depression risk among Korean college students [5]. This study examined 171 family datasets, which included data from dads, mothers, and college students, to uncover major family and individual characteristics linked with depression risk. The study compared the three machine learning models' prediction accuracy: random forest (RF), support vector machine (SVM), and sparse logistic regression (SLR).

Shweta Ware and her colleagues examined if it would be possible to use smartphone data to forecast a person's specific symptoms of depression in a 2020 study [6]. They collaborated to create a set of machine learning models based on characteristics derived from two different kinds of smartphone data: smartphone sensing data and WiFi infrastructure meta-data. Their findings, based on data collected from 182 college students, demonstrated that smartphone data may consistently predict a wide variety of depression symptoms.

The study found that all three machine learning models performed well in prediction, with the RF model beating the others. This model identified five key risk factors for depression: neuroticism, fearful-avoidant attachment, family cohesiveness, self-perceived mental health in college students, and maternal sadness. Furthermore, the logistic regression model multiple risk factors for depression such as: the severity of the mother's pulmonary conditions, the father's illness, and the self-perceived mental health, conscientiousness, and neuroticism of college students.

Laura S. P. Bloomfield observed in 2024 that consumer wearables have promise for measuring sleep habits and perhaps forecasting changes in mental health markers like as stress [7]. However, it remains difficult to establish a strong relationship between physiological measurements of stress-related sleep and an individual's reported stress levels. The study included students from a public institution who participated in the first semester of the Lived Experiences Measured Using Rings Study (LEMURS) from October to December 2022. The connections between predicted sleep metrics and self-reported stress levels were

investigated for 525 individuals using weekly questionnaires and continuous biometric data collecting.

Using mixed-effects regression models, the study tried to find consistent relationships between how we perceive stress levels and various sleep-related metrics, including Average respiratory rate (ARR), heart rate variability (HRV), resting heart rate (RHR), and average nightly total sleep time (TST). Many of these associations remained significant even after adjusting for factors such as gender and semester week. Notably, each additional hour of TST was associated with a 38.3% decrease in the likelihood of experiencing moderate-to-high stress, while a 1 beat per minute increase in RHR was associated with a 3.6% increase in the likelihood of experiencing such stress levels. Additionally, a 1 millisecond increase in HRV corresponded to a 1.2% decrease in the likelihood of moderate-to-high stress, while each additional breath per minute in ARR was linked to a 23.0% increase in the likelihood of experiencing such stress levels. The study also found that participants who did not identify as male reported significantly higher stress levels throughout the study, and stress levels varied across different weeks of the semester. Overall, the findings suggest that sleep data collected from wearable devices could play a valuable role in understanding and predicting stress, which is a significant indicator of the prevalent mental health challenges among college students.

In 2016, Sohrab Saeb did a study that demonstrated the ability of smartphones to identify sadness via passive data collecting from phone sensors [8]. The study sought to duplicate and expand prior studies by using geographic position (GPS) sensors to determine the severity of depressed symptoms.

48 college students were included in the dataset, which was collected over a 10-week period. At the beginning and end of the trial, depression symptom severity was assessed using the Patient Health Questionnaire 9-item (PHQ-9) and GPS phone sensor data. GPS characteristics were calculated throughout the whole research period, including weekdays, weekends, and two-week intervals.

The study replicated prior findings by finding significant connections between GPS characteristics like location variance, circadian mobility and entropy, and PHQ-9 scores (correlations ranging from -0.43 to -0.46 with p-values $<.05$). Interestingly, when GPS characteristics and PHQ-9 scores were derived using weekend data rather than weekday data, higher associations were found. Although the association between baseline PHQ-9 scores and

2-week GPS features reduced with time, substantial correlations with end-of-study scores remained, regardless of the time point utilized for feature derivation.

These findings support earlier research that has shown GPS characteristics to be reliable indicators of how severe the depressed is. The varied advantages of these correlations between weekdays and weekends indicate the impact of weekend dynamics. Moreover, the ability of GPS features to forecast the intensity of depression symptoms for ten weeks in advance highlights their potential as early indicators of depression.

While these studies offer valuable insights into the application of algorithms for detecting depression signs in university students, there remains a need for further research to address challenges such as data privacy, model interpretability, and cultural considerations. By building upon existing work and addressing these challenges, we can advance the development of effective and ethically sound approaches for supporting the mental well-being of university students using machine learning techniques.

2.3 Scope and limitation

The scope of this study encompasses the development and implementation by using machine learning related approaches in detecting depression symptoms in those who study in university. This includes gathering data gathered by various methods such as surveys, medical records, social media, smartphones, and academic performance metrics. This study will explore different algorithms for models and techniques to analyze the data and build predictive models. But, this study has some limitations such as,

1. **Data Availability**: The availability and quality of data from various sources may vary, potentially limiting the scope and generalizability of the findings.
2. **Cultural and Contextual Factors**: What we find from the study might be influenced by cultural and contextual factors [9] related to the university setting and region under investigation, potentially limiting its applicability to other contexts.
3. **Model Performance**: While machine learning models offer promising avenues for mental health assessment, their performance may be influenced by factors such as data imbalance [10], overfitting [11], and the choice of hyper parameters [12].

4. **Ethical Challenges**: Addressing ethical considerations such as data privacy, bias mitigation, and model transparency may pose challenges in practice and require careful consideration and collaboration with relevant stakeholders [13].
5. **Interpretability**: Some machine learning algorithms, particularly complex strategies such as neural networks, may lack interpretability [14], limiting the understanding of how predictions are made and potentially hindering their acceptance and adoption in clinical practice.

2.4 Research Objectives

The following objectives help focus the study and ensure that the research remains aligned with its purpose.

1. Develop models with algorithms to accurately detect depression signs in university students using online survey and social media.
2. Identify key factors for predicting mental health issues among students through feature selection techniques, improving model effectiveness and understanding.
3. Compare different machine learning algorithms (SVM, DT, RF, KNN, NB) to find the most reliable method for detecting depression and anxiety symptoms in university students.
4. Address ethical concerns like privacy and bias in using machine learning for mental health assessment, ensuring responsible and transparent practices.
5. Offer practical recommendations for implementing machine learning-based approaches in universities to support student well-being and academic success.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Research methodology means the plan which is organized and method that investigators employ to carry out a study. It has methods such as the strategies, rules, and procedures employed to collect, preprocess, and interpret data using machine learning models. The methodology ensures that the research is carried out in a structured and scientifically valid manner, enabling the study to produce reliable and credible results.

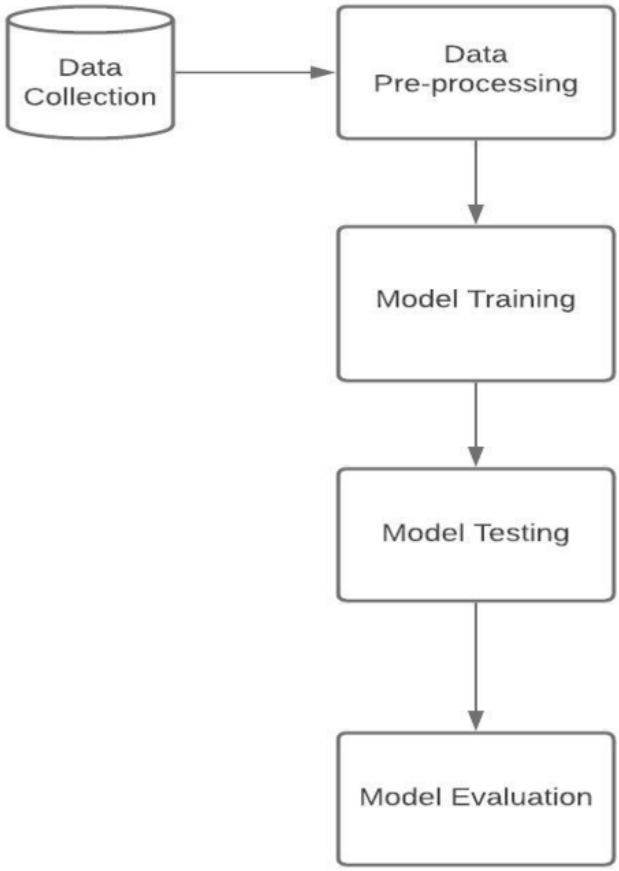

Figure 3.1: Methodology diagram

3.2 Research Subjects

For detecting depression and anxiety symptoms in university students using machine learning, the research subjects are the university students themselves. Here are the key aspects we considered when defining and describing research subjects:

1. **Demographic Characteristics**:
    - **Age**: University students included are within the age limit of 18-35 years.
    - **Academic Level**: Students from different years of study (freshmen, sophomores, juniors, seniors) and different levels of higher education (undergraduate, postgraduate).
    - **Fields of Study**: Including students from various disciplines (e.g., arts, sciences, engineering) to capture a broad range of experiences and stressors.
2. **Geographic and Institutional Characteristics**:
    - **Location**: Students from multiple universities, including both urban and rural settings to account for different stress factors.
    - **Type of Institution**: Public vs. private universities, large research universities vs. smaller colleges, etc.
3. **Psychosocial and Academic Characteristics**:
    - **Academic Performance**: GPA is considered to examine correlations with mental health symptoms [15].

4. **Health and Lifestyle Factors**:
    - **Pre-existing Mental Health Conditions**: Prior diagnosis of depression, anxiety, or other mental health conditions.
    - **Lifestyle**: Information on sleep patterns, diet, exercise, and substance use (e.g., alcohol, drugs, caffeine).

Selection Criteria

1. **Inclusion Criteria**:
    - Enrollment as a current student at the participating universities.
    - Willingness to participate and provide informed consent.
    - Sufficient proficiency in the language of the survey or assessment tools.
2. **Exclusion Criteria**:
    - Incomplete or unreliable responses to surveys and assessments.

Ethical Considerations

- **Informed Consent**: Ensuring students understand the study's objectives,techniques, dangers, and benefits before participating.
- **Confidentiality**: Protecting the privacy of participants by making data anonymous and securely storing personal information.
- **Support Resources**: Providing participants with access to mental health resources and support, especially if the study identifies students with significant symptoms of depression or anxiety.

3.3 Data Collection

Conducting a survey using Google is a convenient and efficient way to collect data from participants. We have used online surveying and available database to gather our data. The survey was restricted to university students. We have collected information of 7022 university students about 15 features for our research. Our survey had two parts. The first part asked question about the independent features and the second part asked question about the target feature. For the first part the questions were,

1. Are you: Male or Female?
2. How old are you?
3. What is your CGPA?
4. How would you describe your recent stress level on a 0 to 5 scale? (0 being not stressed at all to 5 being extremely stressed.)
5. How would you describe your sleep quality? (Poor/ Average/ Good)
6. How physically active are you? (Low/ Moderate/ High)
7. How would you describe your diet quality? (Poor/ Average/ Good)
8. How much social support do you get such as help from family or friends? (Low/ Moderate/ High)
9. What is your relationship status? (Single/ In relationship/ Married)
10. How frequently do you use any substance or drugs? (Never/ occasionally/ frequently)
11. Do your family members have history of depression or mental illness? (Yes/No)
12. Do you have any chronic illness? (Yes/No)

13. How would you describe your financial stress level on a 0 to 5 scale? (0 being not worried about it to 5 being very worried about financial situation)

Using these questions we collected information about our 13 independent features.

We used different set of questions to collect information about the dependent variable depression. It was inspired by beck's depression inventory [16] method.

This depression severity is scored by the survey takers by themselves. The end of the questionnaire has scoring scale which was used to detect depression symptom.

1. 0 I do not feel sad.
   1 I feel sad
   2 I am sad all the time and I can't snap out of it.
   3 I am so sad and unhappy that I can't stand it.
2. 0 I am not particularly discouraged about the future.
   1 I feel discouraged about the future.
   2 I feel I have nothing to look forward to.
   3 I feel the future is hopeless and that things cannot improve.
3. 0 I do not feel like a failure.
   1 I feel I have failed more than the average person.
   2 As I look back on my life, all I can see is a lot of failures.
   3 I feel I am a complete failure as a person.
4. 0 I get as much satisfaction out of things as I used to.
   1 I don't enjoy things the way I used to.
   2 I don't get real satisfaction out of anything anymore.
   3 I am dissatisfied or bored with everything.
5. 0 I don't feel particularly guilty
   1 I feel guilty a good part of the time.
   2 I feel quite guilty most of the time.
   3 I feel guilty all of the time.
6. 0 I don't feel I am being punished.
   1 I feel I may be punished.
   2 I expect to be punished.
   3 I feel I am being punished.

Figure 3.3.1: Questions 1-6

7. 0 I don't feel disappointed in myself.

   1 I am disappointed in myself.

   2 I am disgusted with myself.

   3 I hate myself.

8. 0 I don't feel I am any worse than anybody else.

   1 I am critical of myself for my weaknesses or mistakes.

   2 I blame myself all the time for my faults.

   3 I blame myself for everything bad that happens.

9. 0 I don't have any thoughts of killing myself.

   1 I have thoughts of killing myself, but I would not carry them out.

   2 I would like to kill myself.

   3 I would kill myself if I had the chance.

10. 0 I don't cry any more than usual.

    1 I cry more now than I used to.

    2 I cry all the time now.

    3 I used to be able to cry, but now I can't cry even though I want to.

11. 0 I am no more irritated by things than I ever was.

    1 I am slightly more irritated now than usual.

    2 I am quite annoyed or irritated a good deal of the time.

    3 I feel irritated all the time.

12. 0 I have not lost interest in other people.

    1 I am less interested in other people than I used to be.

    2 I have lost most of my interest in other people.

    3 I have lost all of my interest in other people.

Figure 3.3.2: Questions 7-12

13. 0 I make decisions about as well as I ever could.

    1 I put off making decisions more than I used to.

    2 I have greater difficulty in making decisions more than I used to.

    3 I can't make decisions at all anymore.

14. 0 I don't feel that I look any worse than I used to.

    1 I am worried that I am looking old or unattractive.

    2 I feel there are permanent changes in my appearance that make me look unattractive

    3 I believe that I look ugly.

15. 0 I can work about as well as before.

    1 It takes an extra effort to get started at doing something.

    2 I have to push myself very hard to do anything.

    3 I can't do any work at all.

16. 0 I can sleep as well as usual.

    1 I don't sleep as well as I used to.

    2 I wake up 1-2 hours earlier than usual and find it hard to get back to sleep.

    3 I wake up several hours earlier than I used to and cannot get back to sleep.

17. 0 I don't get more tired than usual.

    1 I get tired more easily than I used to.

    2 I get tired from doing almost anything.

    3 I am too tired to do anything.

18. 0 My appetite is no worse than usual.

    1 My appetite is not as good as it used to be.

    2 My appetite is much worse now.

    3 I have no appetite at all anymore.

Figure 3.3.3: Questions 13-18

19. 0 I haven't lost much weight, if any, lately.

    1 I have lost more than five pounds.

    2 I have lost more than ten pounds.

    3 I have lost more than fifteen pounds.

20. 0 I am no more worried about my health than usual.

    1 I am worried about physical problems like aches, pains, upset stomach, or constipation.

    2 I am very worried about physical problems and it's hard to think of much else.

    3 I am so worried about my physical problems that I cannot think of anything else.

Figure 3.3.4: Questions 19-20

After completing the questionnaire, students were asked to calculate the score for each of the twenty queries by calculating the number of choices they selected. The highest possible score for the entire test is sixty. This indicates that an individual marked highest number for all questions. The lowest rating for question is 0; hence the lowest obtainable value for the exam is also 0.

A score of 0-25 shows that the student is not depressed. If the score exceeds 25, the student has some form of depression, ranging from moderate to severe.

3.4 Data Preprocessing

Data Preprocessing is a criterion which is important in the data analysis and machine learning pipelines. It entails changing and reforming raw data into a clean, readable structure before putting it into the model. Here are the key steps used by us for our data preprocessing:

1. **Data Cleaning**:
    - **Manage Missing Values**: We treated missing data for cgpa and substance usage by either eliminating the records with values not present or filling the gaps using mean and median.

2. **Data Integration**:
    - Combining data from different sources and ensuring consistency. This step involved dealing with redundancies, and resolving data conflicts.

3. **Data Transformation**:
    - **Normalization/Standardization**: Scaling number characteristics to a certain range (for example, 0 to 1) or standardizing them with standard deviation. Ordinal values [17] with relevant orders are converted into values on a 0 to 2 or 0 to 5 scale. For example, sleep quality ratings of poor, average, and good are converted to 0/1/2 numbers.

    - **Encoding Categorical Variables**: Converting categorical data to numerical representation via tools such as one-hot encoding [18], label encoding, or binary encoding. Such as, information on males and females has been converted to binary, with true indicating male and false indicating female.

4. **Data Partitioning**:
    • The dataset was spitted and maintained into test sets and training sets to appropriately assess the performance and avoid overfitting.

Each of these steps is crucial for ensuring that the data is in the best possible shape for analysis and modeling, thereby improving the performance and reliability of the results.

3.5 Dataset

Datasets are fundamental units in data analysis, machine learning, and various data-driven applications. The dataset has 14 features and they are 'Age', 'CGPA', 'Stress_level', 'Depression_Score', 'Sleep_Quality', 'Physical_Activity', 'Diet_Quality', 'Social_Support', 'Relationship_Status', 'Substance_Use', 'Family_History', 'Chronic_Illness', 'Financial_Stress', 'Gender_Male'. Here, 'Depression_Score' is the target dependent feature while the other features are independent.

Features can be of different types such as,

- **Numerical**: Quantitative values (e.g., age, income).

- **Categorical**: Qualitative values (e.g., gender, color).
- **Ordinal**: Values with a meaningful order (e.g., ratings, education levels).
- **Temporal**: Time-related data (e.g., dates, timestamps).
- **Text**: Free-form text data (e.g., reviews, comments).

In our dataset, 'Age', 'CGPA' are numerical features. 'Gender_Male' is a categorical feature. Rests of the features are ordinal features.
In our dataset, out of 7022 students, 1727 students have some level of depression while 5295 students don't have any depression.

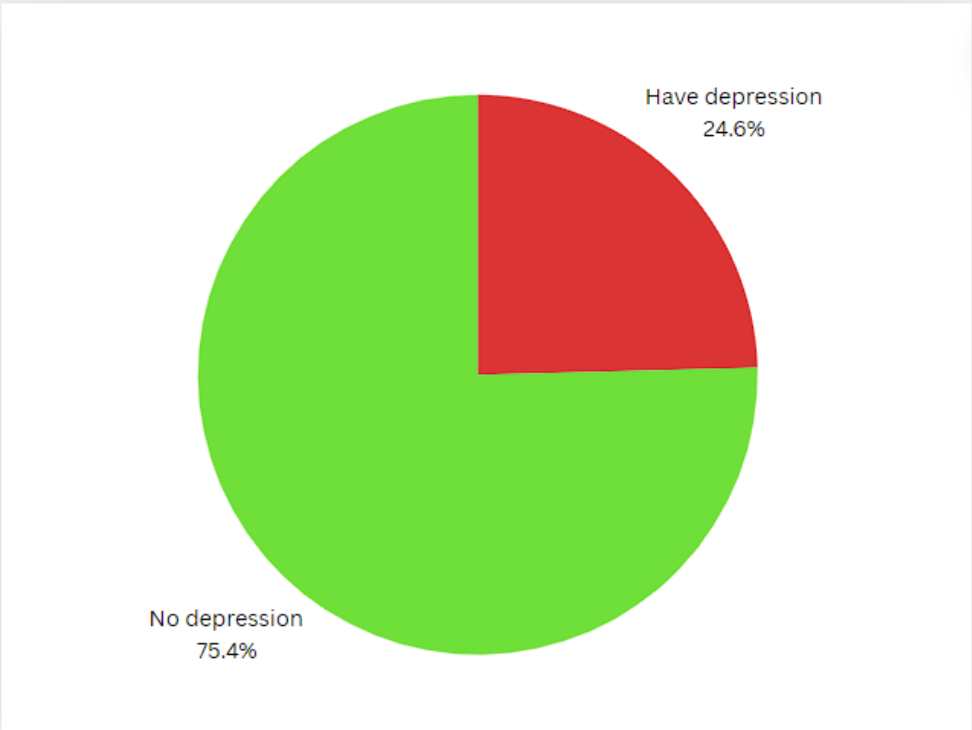

Figure 3.5: Pie chart showing percentage of students who have depression

3.6 Statistical Analysis

Statistical analysis meaning the process of gathering, arranging, purifying, analyzing, and showcasing data to reveal meaningful patterns, trends,and knowledge[19]. It encompasses a wide range of techniques and methods designed to extract valuable information from datasets, enabling researchers to make informed decisions and draw valid conclusions.

The process of statistical analysis typically begins with data collection, where necessary related information is collected from various places such as surveys, experiments, or observational studies. Once we collected the data, it underwent thorough cleaning and preprocessing to remove any inconsistencies, errors, or missing values that could distort the analysis.

We have 3547 male students and 3475 female students in the dataset. The students in the dataset aged from 18 years to 35 years.

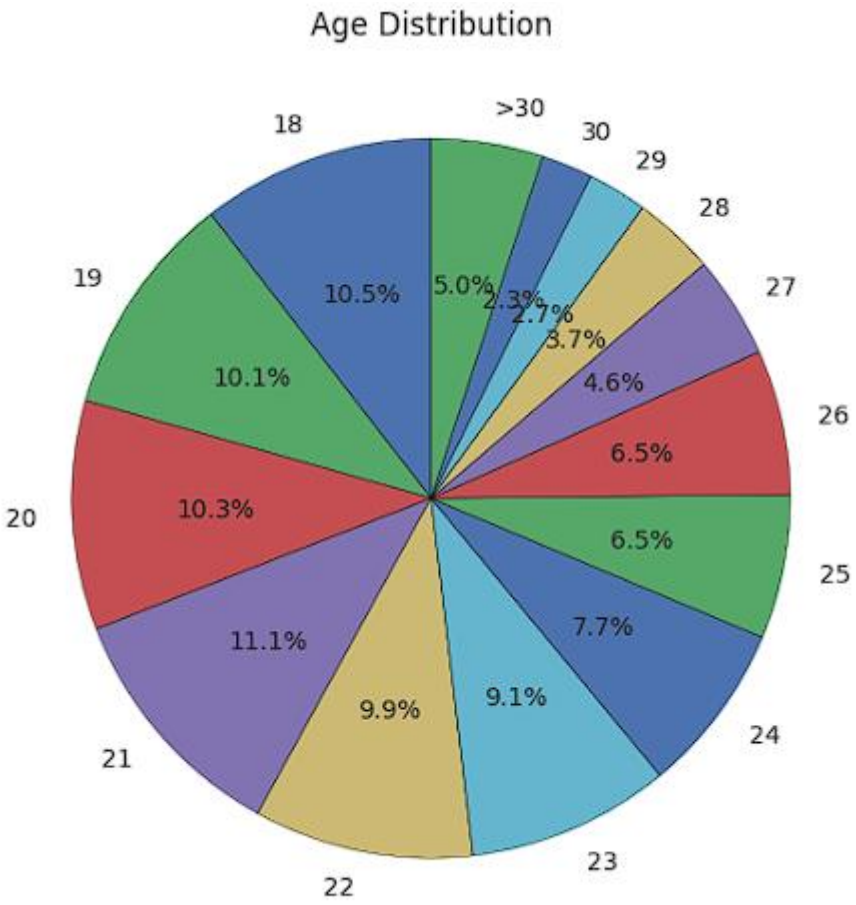

Figure 3.6.1: Pie chart showing age distribution of students

Here we can see that the largest age group is 21 years old with 11.1%. We can also see that more than three quarters of the students are younger than 26 years old while students older than 30 are only 5% of the students.

This distribution of age groups underscores the youthful demographic composition of the student body surveyed, with a predominant concentration in the early to mid-twenties age range. Such insights into the age distribution of the student population can offer valuable context for understanding the demographic characteristics of the sample and may inform tailored approaches to addressing mental health issues and support needs within this cohort.

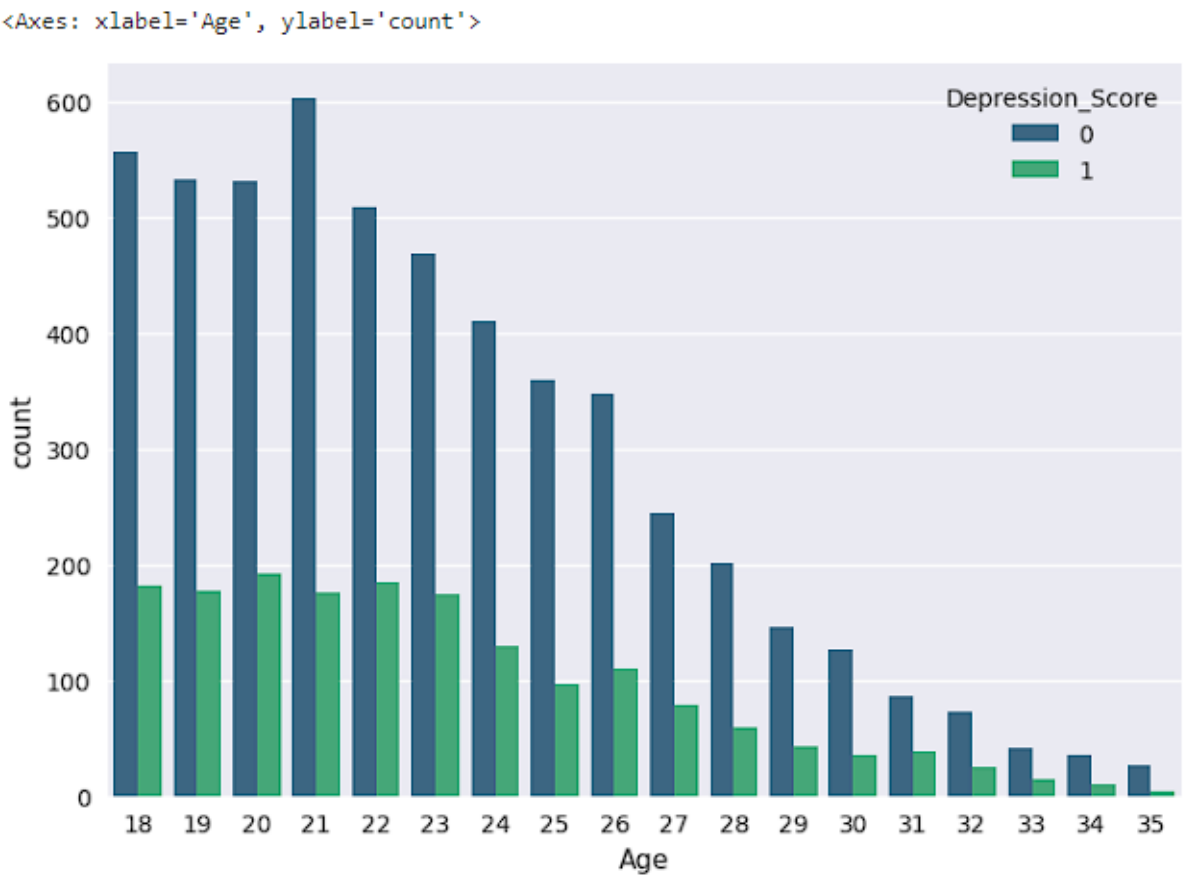

Figure 3.6.2: Bar chart showing different age groups affected by depression.

From the chart we can see that, even though 21 years old age group is the largest, 20 years olds have the highest number of depressed students. Students from 18 to 23 years olds have similar percentage of depressed students while the percentage of depressed students varies for students older than 23 years old.

The data from the chart indicates an intriguing trend regarding the prevalence of depression across different age groups among students. While the 21-year-old age group emerges as the

largest demographic segment, it is notable that the highest number of depressed students is observed among 20-year-olds. This suggests a potential peak in depressive symptoms among individuals at this particular age, despite being outnumbered by 21-year-olds in terms of overall population size. Moreover, the analysis reveals a relatively consistent percentage of depressed students among those aged between 18 and 23 years old, indicating a similar prevalence of depression within this range. However, beyond the age of 23, the percentage of depressed students appears to vary more significantly, suggesting potential shifts in mental health dynamics as students' progress into their mid-to-late twenties and beyond. Further exploration of these variations in depression prevalence across different age groups could provide valuable insights into the factors influencing mental health outcomes during different stages of young adulthood, guiding the development of targeted interventions and support services tailored to the specific needs of students within each age bracket.

3.7 Model Training and Evaluation

We have trained multiple machine learning models by using algorithms such as including Support Vector Machines (SVM),Decision Trees (DT),Random Forests (RF), k-Nearest Neighbors (KNN), and Naive Bayes (NB), using the cleaned and preprocessed data.

3.7.1 Model 1: Decision Trees (DT)

The Decision Tree algorithm is a very widely used machine learning algorithms, using a tree-like structural frame and many combined methods to solve specific issues [20]. It falls within the realm of supervised learning, adept at managing both classification and regression tasks.

A Decision Tree is a powerful machine learning algorithm that can be effectively utilized to detect depression from a dataset. This algorithm works by partitioning the data into sets considering by value of input features, constructing a tree-like model of decisions that can classify whether an individual is likely experiencing depression.

In the context of detecting depression, the dataset might include various features such as patient responses to psychological questionnaires, demographic information, lifestyle factors, and medical history. The Decision Tree structure begins with a root node, which represents the entire dataset. This node is then split into subsets based on the values of selected features,

creating branches that divide into internal nodes. Each internal node represents a choice point according to a specific feature, while the leaf nodes represent the final classification—whether the individual is likely to be experiencing depression or not.

The method of creating a decision tree for identifying depression begins with picking the feature that best separates the data, which may be found using metrics such as Gini impurity or information gain. For example, a question regarding sleep habits or emotions of sorrow may be the first characteristic used to segment the data. This splitting procedure is repeated for each subgroup until specified ending requirements are satisfied, such as a maximum tree depth or a minimum number of samples per node, to ensure that the model does not become unduly complicated.

Pruning strategies are used to avoid overfitting, a typical problem with decision trees. Pruning eliminates branches that do not significantly add to the model's prediction capacity, increasing its ability to generalize to new, previously unknown data. This phase is critical for ensuring that the model stays resilient and accurate in recognizing depression in the dataset.

Decision trees have various advantages in this scenario. They are easily interpretable, helping healthcare practitioners to comprehend the decision-making process and identify significant elements connected with depression. Furthermore, they do not require data normalization or scaling and can handle both linear and nonlinear connections between characteristics and depression risk.

However, decision trees can have certain limitations. They may easily overfit training data if not properly pruned, and they are sensitive to minor changes in the dataset, resulting in various tree architectures. Furthermore, if the dataset is skewed, with more cases of non-depression than depression, the tree may become biased toward the more common class. We used Gini index [21] as selection measure.

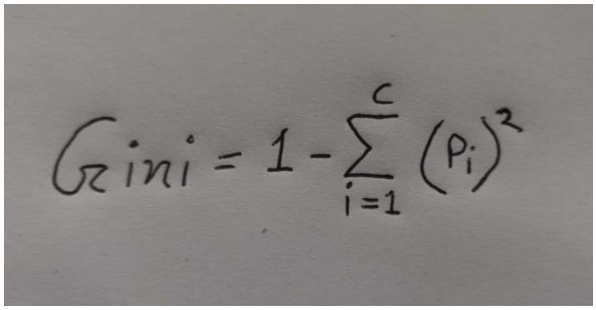

Figure 3.7.1.1: Gini index

Using Gini index and max-depth of 3 we got precision score of 0.7558 which is supposed to mean that this model could correctly detect the depression status of 75.58% of the total students.

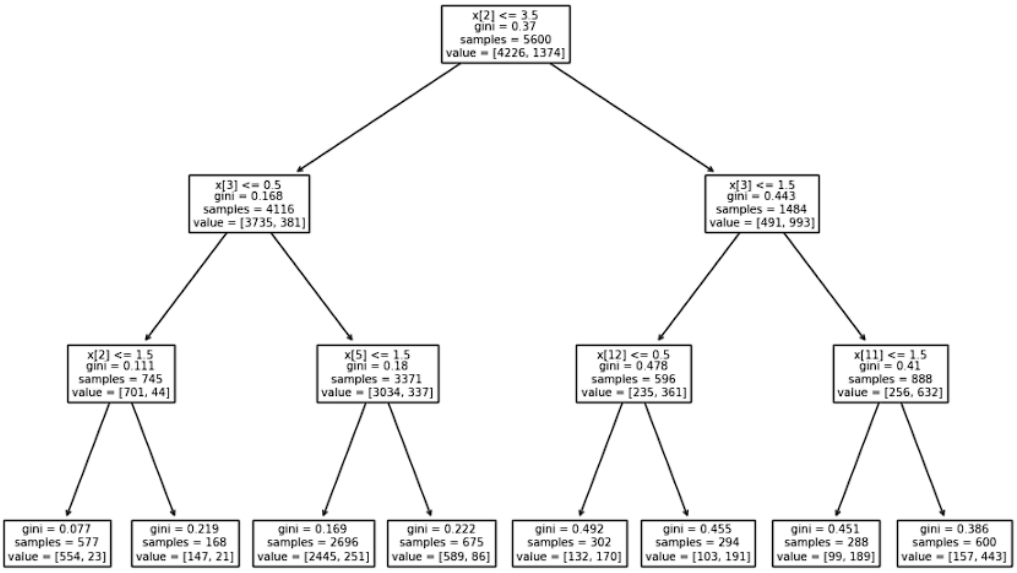

Figure 3.7.1.2: Visualization of decision tree classifier using gini index and max depth 3

3.7.2 Model 2: k-Nearest Neighbors (KNN)

The most simple approaches in machine learning is the k Nearest Neighbors (kNN) algorithm. It is a non-parametric approach used for classification and regression applications [22]. The k-Nearest Neighbors (k-NN) algorithm is an uncomplicated yet effective supervised machine algorithm that can be applied to detect depression from a dataset. This

algorithm relies heavily on the principle of similarity to guess a data point's class by understanding the classes of its closest neighbors in the feature space. In the context of detecting depression, the dataset might include various features such as patient responses to psychological questionnaires, demographic information, lifestyle habits, and medical history. The k-NN algorithm classifies an individual as likely experiencing depression or not by examining the features of other individuals in the dataset who have already been classified.

To implement k-NN for depression detection, several steps are involved. First, relevant features are selected that can help identify depression. These features might include responses to specific questions about mood, sleep patterns, energy levels, and other behavioral indicators. The algorithm then calculates the space between the data place (individual) in question and all other places in the dataset using a chosen distance metric, such as Euclidean distance, which measures similarity based on the responses to depression-related questions. The parameter kkk, representing the number of nearest neighbors to consider, is selected. Choosing an appropriate value for kkk is crucial, as a small kkk might make the model sensitive to noise, while a large kkk might smooth out important distinctions

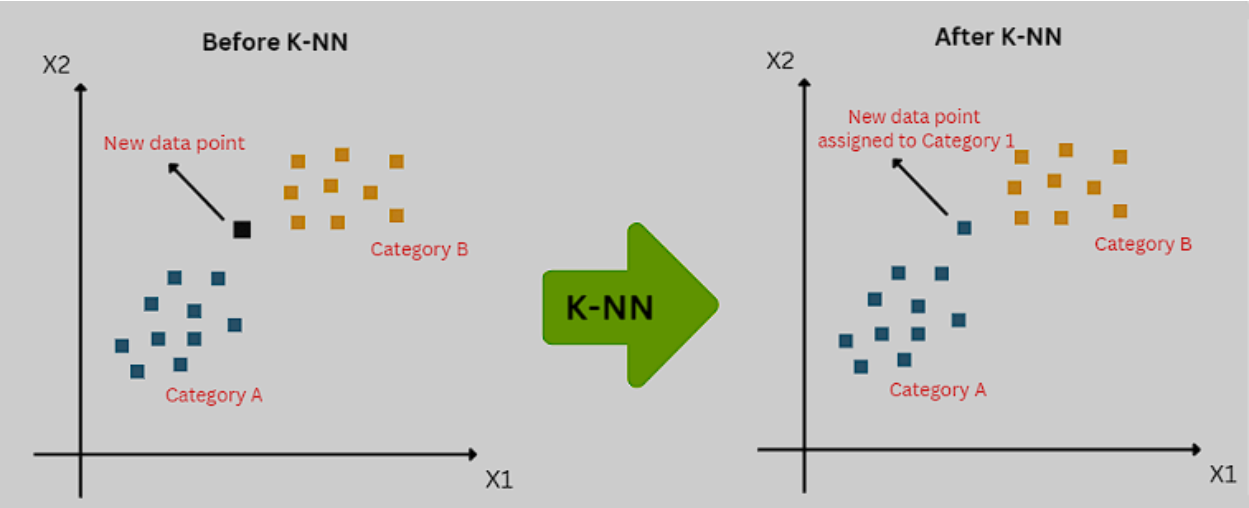

Figure 3.7.2: Visualization of K-NN algorithm process

When using k-Nearest Neighbors (k-NN) to diagnose depression, selecting proper characteristics is important to the algorithm's efficacy. Clinical measures such as the frequency of depressive symptoms, responses to standardized mental health questionnaires, patient demographics (age, gender, socioeconomic status), lifestyle factors (sleep patterns, physical activity), and medical history are all typically taken into account. By carefully picking and designing these characteristics, the k-NN algorithm may better detect depression-related patterns.

One of the primary benefits of utilizing k-NN for depression detection is its simplicity and convenience of implementation. In contrast to more complicated models such as neural networks, the method does not require lengthy training. Instead, k-NN saves the full dataset and classifies new cases by comparing them to previously stored examples. This makes it particularly useful for tiny datasets with low processing resources. Furthermore, k-NN can handle both linear and non-linear correlations between characteristics and the target variable, which is useful in the context of mental health, where symptoms and their interactions can be complex and multidimensional.

However, adopting k-NN for this purpose comes with its own set of obstacles. One key difficulty is computational inefficiency with big datasets, as the method must calculate the distance between the query instance and each instance in the dataset, which can be time-consuming. Furthermore, the performance of k-NN is significantly dependent on the distance measure used and the value of k. Experimentation and cross-validation are frequently used to balance bias and variation while determining the optimal k. A low k may result in overfitting, in which the model catches noise in the training data, whereas a high k may result in underfitting, in which the model becomes overly generic and misses significant patterns.

Another problem is how to handle skewed datasets, which are prevalent in depression diagnosis. Typically, there are fewer positive instances (individuals with depression) than negative cases (individuals without depression). This imbalance may bias the k-NN algorithm to predict the majority class. To address this, techniques such as re-sampling the dataset (over-sampling the minority class or under-sampling the majority class), weighted voting (in which the influence of each neighbor is weighted by distance), and adjusting the decision threshold can be used to improve model performance.

Eager learners [23] and lazy learning [24] are utilized in this model. Using the kNN algorithm and neighbors number of 5 we got the highest precision score of 0.8036 which is supposed to mean that the algorithm could accurately detect the depression status of 80.36% of the total students.

3.7.3 Model 3: Support Vector Machines (SVM)

Support Vector Machines (SVMs) are algorithms in machine learning used for classification and regression applications. SVMs are well-known for their resilience; they excel in classifying data points and detecting outliers [25]. Support Vector Machines (SVMs) are a sophisticated supervised machine learning method that may successfully identify sadness in a dataset. SVM works by determining the ideal hyperplane for dividing the data into classes. In the context of identifying depression, the dataset may comprise variables such as psychological evaluation results, demographic data, lifestyle factors, and medical history. SVM uses these factors to identify between those who are likely to be depressed and those who are not.

The SVM method works by changing the input data into a high-dimensional space and finding for the hyperplane that maximizes the distance between different classes. The distance between the hyperplane and the support vectors—the nearest data points in each class—defines the limit. These support vectors are critical for creating the hyperplane, and SVM uses them to generate the model. When identifying depression, SVM may utilize data such as mood ratings, sleep habits, and activity levels to establish this distinction.

One of the primary benefits of employing SVM for depression identification is its capacity to work in dimension high areas and manages both linear and non-linear connections between different features. SVM may handle complicated associations in data by transforming it into a higher-dimensional space with a linear separator using kernel functions like the polynomial kernel, radial basis function (RBF), or sigmoid kernel. This is especially beneficial in mental health, where the link between symptoms and depression can be complex and nonlinear.

However, applying SVM for depression diagnosis presents obstacles. One major challenge is selecting an appropriate kernel and tweaking hyperparameters such as the parameter C and kernel parameters. These decisions have a significant impact on model performance and need careful adjustment, which is frequently accomplished through cross-validation. Additionally,

SVMs may be computationally costly, particularly with big datasets, because the algorithm must solve a complicated optimization issue. This can be minimized by employing techniques like stochastic gradient descent or utilizing significant computational resources.

Another issue is the possibility of overfitting, which occurs when the model is very sophisticated in comparison to the quantity of data given. Regularization methods and thorough cross-validation can assist to mitigate this risk. Furthermore, SVMs are sensitive to the dataset's balance. In cases when there is a considerable class imbalance (more people without depression than with depression), measures such as class weighting, synthetic data creation, or anomaly detection procedures may be required to guarantee the model performs well across both classes.

Support Vector Machines provide a strong foundation for identifying depression in datasets by utilizing their abilities to handle high-dimensional data and complicated interactions via kernel functions. Despite the problems of parameter tweaking, computing needs, and the possibility of overfitting, SVMs' value rests in their flexibility and efficacy, making them an important tool in the field of mental health diagnoses. With correct implementation and tweaking, SVMs may produce accurate and trustworthy predictions, assisting in the early identification and treatment of depression.

The purpose of SVMs is to discover the hyperplane with the biggest possible margin between support vectors inside the dataset [27]. It employs the kernel method to improve classifier accuracy and solve nonlinear separation problems.

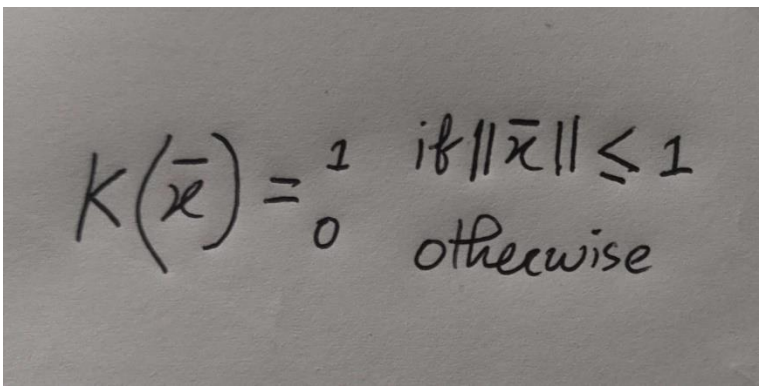

Figure 3.7.3.1: Kernal Function

When the data can be separated linearly, a linear kernel is utilized [28]. The linear kernel has one noteworthy advantage which is its interpretability. The resultant linear decision boundary is logical and easy to understand, allowing for insights into how different characteristics contribute to depression identification. Furthermore, the linear kernel is computationally efficient, making it ideal for datasets of significant size. Unlike more complicated kernels such as polynomial or (RBF) kernels, the linear kernel does not need translating data into higher-dimensional regions, hence lowering processing complexity. The graphic below depicts a linear kernel.

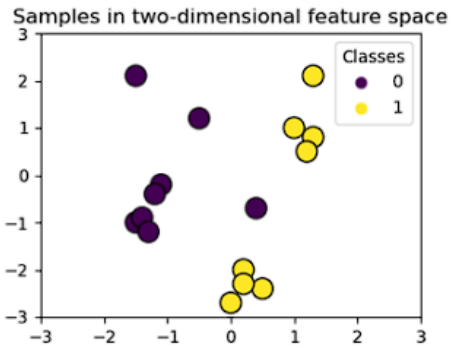

Figure 3.7.3.2: Linear Kernel

We have also used polynomial kernel. The polynomial kernel is a kernel function used in Support Vector Machines (SVM) and other kernelized models to handle non-linear input by transferring it to a higher-dimensional space.

At the end, we got better result using the linear kernel and got an accuracy of 83.37%.

3.7.4 Model 4: Random Forests (RF)

A learning method called Random Forest with two separate variations—one designed for classification problems and the other for regression [29]. In terms of diagnosing depression from a dataset, the Random Forests algorithm appears as a powerful and adaptable tool. Random Forests are an ensemble learning approach that is made up of numerous decision trees. Each decision tree in the forest is created individually, using a portion of the dataset and a random collection of characteristics. This randomization adds variation to the individual trees, reducing overfitting and improving the ensemble's overall predictive accuracy.

Random Forests are ideal for detecting sadness because of their capacity to handle high-dimensional data with complicated feature relationships. The algorithm can efficiently evaluate a wide range of variables usually found in mental health datasets, such as demographic information, behavioral patterns, and clinical assessments. Random Forests, which aggregate forecasts from several trees, may capture varied patterns in data, resulting in strong and trustworthy predictions about an individual's chance of having depression.
One major primary benefits for Random Forests is their resistance to overfitting, which is a typical problem in machine learning models, especially when working with tiny or noisy datasets. Random Forests reduce the danger of overfitting by combining abilities for predictions from many trees trained by using different subsets of data, resulting in many more generalizable conclusions. Furthermore, Random Forests provide built-in processes for determining feature relevance, allowing researchers and doctors to find the most relevant elements in depression identification.

Furthermore, Random Forests are highly interpretable, giving stakeholders insight into the decision-making process. Unlike more sophisticated models such as neural networks, Random Forests give clear explanations for their predictions, making them especially useful in therapeutic contexts where interpretability is critical. This interpretability encourages confidence and allows machine learning practitioners and mental health subject specialists to collaborate more effectively.

Random Forests, while their many advantages, have certain limits. They may struggle to capture nuanced nonlinear correlations between characteristics and depression as compared to more adaptable methods such as deep learning algorithms. Furthermore, the computing resources needed to train and evaluate Random Forests can be significant, especially for big datasets with many characteristics. However, advances in parallel processing and optimization approaches have helped to address some of these issues, making Random Forests a feasible choice for depression identification in research and clinical practice.

Random Forests are an effective and interpretable tool for detecting sadness in a dataset. Their capacity to handle high-dimensional data, reduce overfitting, and give insights on feature relevance makes them ideal for the difficult task of mental health diagnosis. Random

Forests, which capitalize on the capabilities of ensemble learning, are a powerful tool for academics and therapists looking to better our knowledge and diagnosis of depression.

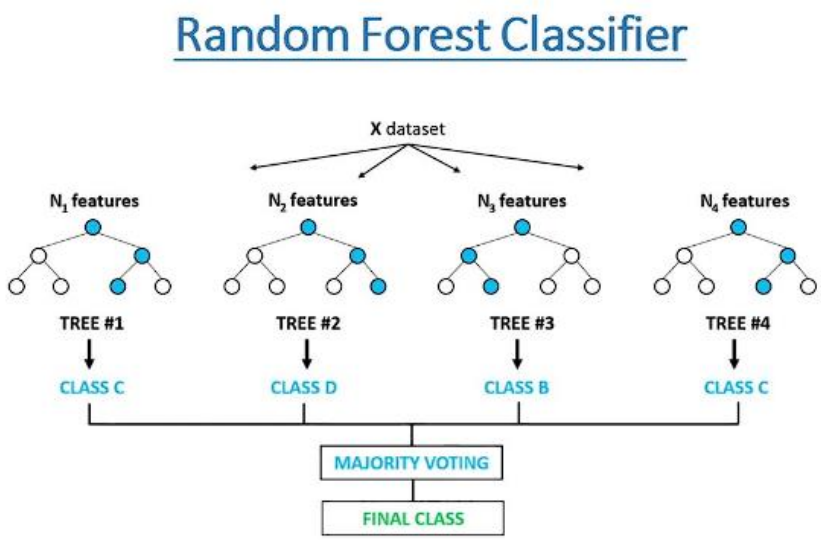

Figure 3.7.4: Random forest classifier process

Number related to decision trees in a Random Forests algorithm is an important parameter that directly affects the model's performance and behavior. Random Forests are made up of an group of decision trees, each of which is separately trained on a random selection of training data and characteristics. The number of decision trees in the ensemble is commonly set by the user as a hyperparameter, and it has a lot of influence on the model's predicted accuracy, stability, and computing efficiency.

Choosing the right number for decision trees in a Random Forests model requires a trade-off between bias and variance. A higher number of trees often reduces variance since the ensemble averages out the predictions of individual trees, resulting in a smoother decision boundary and improved generalization to new data. However, increasing the number of trees has a computational cost, since each additional tree demands more training time and memory resources.

On the other side, having too few decision trees might result in significant bias, causing underfitting and bad performance while testing using both the training datasets and test datasets. In such instances, the Random Forests model may fail to detect complicated patterns in the data, resulting in limited prediction value. As a result, it is critical to choose an appropriate number of decision trees to strike a balance between bias and variance, improving the model's performance while avoiding overfitting.

The random forest model accuracy score with 10 decision-trees is 82.92%.

3.7.5 Model 5: Naive Bayes (NB)

In machine learning, Naïve Bayes classification is a simple yet effective technique for classification tasks [30]. The Naive Bayes algorithm offers a simple yet effective technique to categorization. Naive Bayes uses probability theory concepts to determine the chance of an individual having depression based on their characteristics, making it ideal for situations where understanding the underlying probabilities of various outcomes is critical. Because of its simplicity, efficiency, and interpretability, this algorithm has gained popularity in a variety of fields, including text categorization, sentiment analysis, and medical diagnostics.

When used to depression detection, Naive Bayes makes the assumption that features are given class label which are not same, which might not be necessarily be the truth in real life settings. Despite this simplistic assumption, Naive Bayes can nevertheless produce competitive results by properly modeling the conditional probabilities of features based on the class label. By evaluating data patterns and calculating the likelihood of depression based on observable traits, Naive Bayes can give useful insights into the underlying causes of depression, allowing for more informed patient treatment decisions.

One of the most noticeable advantages of Naive Bayes is its computational efficiency, especially when working with huge datasets. The technique uses minimum training data to estimate classification probabilities, making it ideal for applications with low computing resources. Furthermore, Naive Bayes is resistant to irrelevant features and can manage missing values in the dataset, reducing the preparation processes needed for analysis.

Despite its simplicity and effectiveness, Naive Bayes may struggle to identify complicated correlations between characteristics and depression, especially when factors interact significantly. The assumption of feature independence may limit the algorithm's capacity to detect minor connections in the data, potentially resulting in inferior performance in some cases. Furthermore, Naive Bayes is sensitive to the quality of the training data and may yield biased results if the dataset is unbalanced or has noisy observations.

Naive Bayes is a useful method for detecting depression in datasets because it is simple, efficient, and easy to understand. While Naive Bayes may not capture all of the data's intricacies, it can give useful insights into the elements that contribute to depression and help doctors make educated patient care decisions. Naive Bayes, which uses probabilistic principles, is a helpful addition to the repertory of machine learning algorithms in mental health diagnosis.

We used Gaussian Naïve Bayes for this model. When dealing with continuous attribute values, we assume that the values associated with each class follow a Gaussian or normal distribution [31]. Gaussian Naïve Bayes is a probabilistic technique that assumes features have a Gaussian distribution. This Naïve Bayes method works well with continuous-valued features, making it ideal for datasets with numerical variables like age, mood ratings, or physiological measures. Gaussian Naïve Bayes models each class's feature distribution as a normal distribution, calculating the likelihood of a data point belonging to each class based on observed feature values.

Gaussian Naïve Bayes, like the normal Naïve Bayes method, assumes feature independence based on the class label for detecting depression. Unlike discrete or categorical feature distributions, Gaussian Naïve Bayes models continuous features as Gaussian distributions with mean and variance parameters for each class. The approach can calculate the probability density function (PDF) of a data point's feature values under each class's distribution after estimating the mean and difference of each feature for every class from the training data.

Gaussian Naïve Bayes is a simple and computationally efficient approach for detecting depression. The approach uses only little sample of training data when estimating the parameters of Gaussian distributions, making it ideal for applications with low computing resources. Gaussian Naïve Bayes can manage missing values and irrelevant characteristics, simplifying preprocessing and analysis.

Gaussian Naïve Bayes, despite its simplicity, may achieve competitive performance in depression detection tasks when paired with proper feature selection and preprocessing approaches. Gaussian Naïve Bayes uses Gaussian distributions to capture patterns in data and accurately predict depression status.

Gaussian Naïve Bayes, like ordinary Naïve Bayes, may have difficulty capturing complicated correlations between characteristics and depression, especially when interactions are strong. Furthermore, the assumption of feature independence may restrict the algorithm's capacity to detect subtle connections in the data, potentially resulting in inferior performance in some cases.

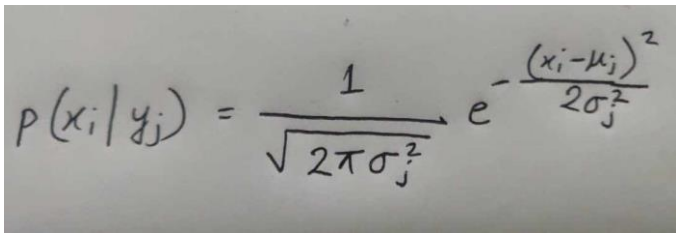

Figure 3.7.5: Gaussian Naïve Bayes algorithm

Using Gaussian Naïve Bayes algorithm the Naive Bayes model accuracy score was 0.8372 which mean it successfully detected depression status of 83.72% of the students.

Table 3.7: Parameter Usage of algorithms

Algorithms	Parameters
Model-1, Decision Tree	Criterion = ‘ entropy’, min_samples_split = 2, max_depth=3
Model-2, KNN	n_neighbors = 3
Model-3, SVM	C=1.0, kernel='linear'
Model-4, Random Forest	n_estimators=100, random_state=0
Model-5, Naïve Bayes Classifier	Random_state = 0, classifier = GaussianNB

Here, the Parameter Usage table outlines the specific parameters and their corresponding values utilized for different machine learning algorithms. These parameters were selected to optimize performance and achieve optimal outcomes.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

The experimental results and subsequent discussion presented in this report serve as a critical component in evaluating the efficacy and performance of the implemented methodologies. Through rigorous experimentation, we aim to shed light on the effectiveness of various algorithms and techniques employed in addressing the problem at hand. This section delves into the empirical findings derived from the conducted experiments, providing insights into the behavior, strengths, and limitations of the proposed approaches. By analyzing the experimental results in detail and engaging in constructive discussion, we endeavor to glean valuable insights that help us increase our comprehensive understanding of the problem domain and inform potential way for further research and refinement.

4.2 Results

We applied different algorithms, such as KNN (k Nearest Neighbor), Naive Bayes classifier, Support Vector Machine (SVM), Decision Tree, and Random Forest, for purposes of correctly identifying the signs. Upon analyzing the accuracy table, we observed that for a data usage rate of 30%, the Naïve Bayes model achieved the highest accuracy of 83.72%.

Table 4.2: Accuracy data of algorithms

Data usage rate	**Algorithms**				
	Decision Tree	***KNN***	***SVM***	***Random Forest***	***Naive Bayes***
20%	75.58%	80.36%	82.99%	82.92%	83.49%
30%	74.46%	78.55%	82.39%	81.87%	83.72%
40%	73.51%	80.17%	82.66%	81.81%	81.42%
50%	74.08%	80.18%	83.37%	81.86%	82.57%

We employed a specific technique for splitting the training and testing data, using test data usage rates ranging from 20% to 50%. This means that when the test size is 20%, the training size is automatically 80%, and vice versa when the test size is 50%, the training size is 50%. This approach was chosen to determine the optimal data usage percentage for our algorithms. Regarding data quality, it is important to assess whether the dataset contains vulnerabilities. If the dataset is highly vulnerable, it will not perform well with a smaller amount of training data.

4.3 Discussion

The Naïve Bayes model was the most accurate among the models we tested. We will test several levels of accuracy to determine how excellent and sensitive this model is. We started by checking for null accuracy and comparing the model's accuracy.

4.3.1 Null accuracy

Null accuracy serves as a baseline measure in classification problems to assess a model's effectiveness. It represents the accuracy achievable by consistently predicting the most frequent class in the dataset, without involving any model training. By checking null accuracy we got the value 75.89% and Naïve Bayes model accuracy is 83.72%. So, Naïve Bayes model is performing better than the null accuracy.

4.3.2 ROC curves

ROC curves explain the comparable difference between true positive rates and false positive rates for a model which predicts with various probability thresholds.

To put it simply, we can categorize it using a prediction model and various probability criteria. In this instance, we'll get both genuine and erroneous positive predictions. After creating a Confusion Matrix and computing the true and false threshold values, each threshold requires its own confusion matrix. Finally, there will be a large number of confusion matrices. Instead, Receiver Operator Characteristics (ROC) can be calculated to offer a summary of the False Positive Rate (FPR) and True Positive Rate (TPR). These will be planned on the X and Y axes, respectively.

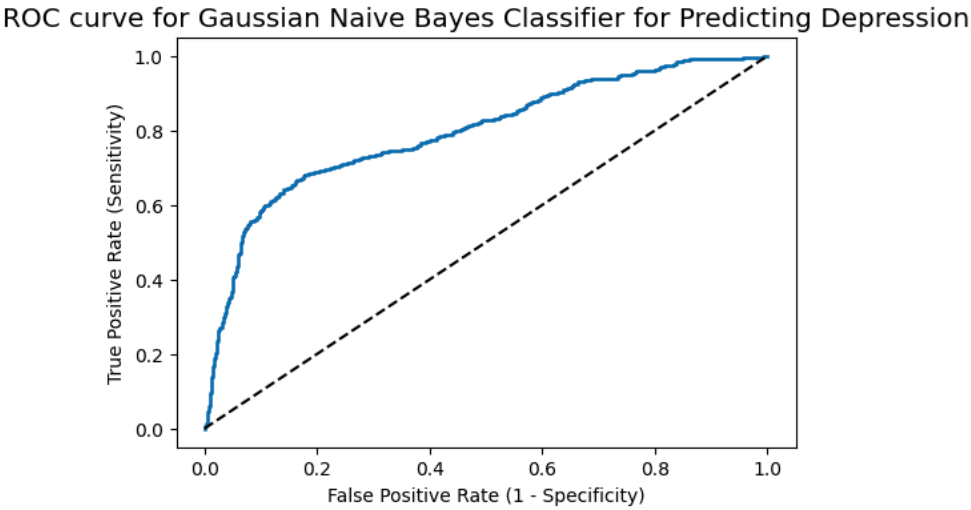

Figure 4.3.2.1: ROC curve for Gaussian Naïve Bayes Classifier.

From the figure we can see Naïve Bayes model is doing better a lot better than no skill prediction. We can check for the ROC curve of other models to compare.

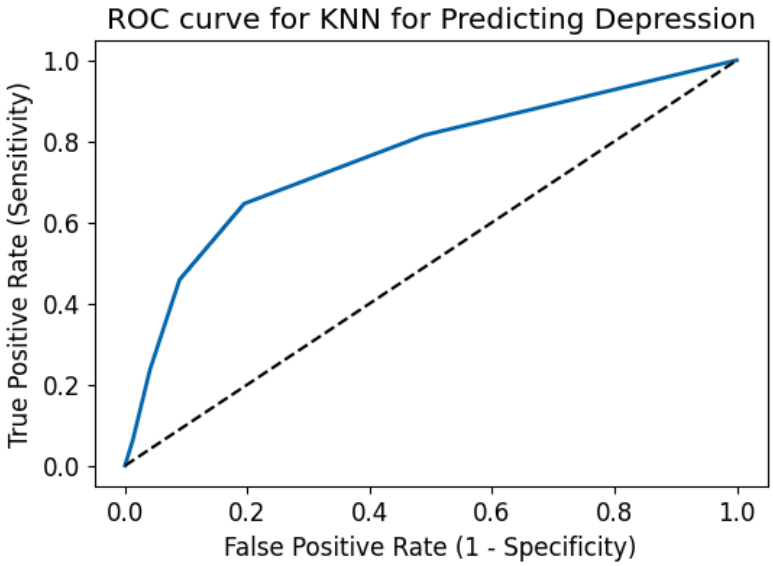

Figure 4.3.2.2: ROC curve of KNN

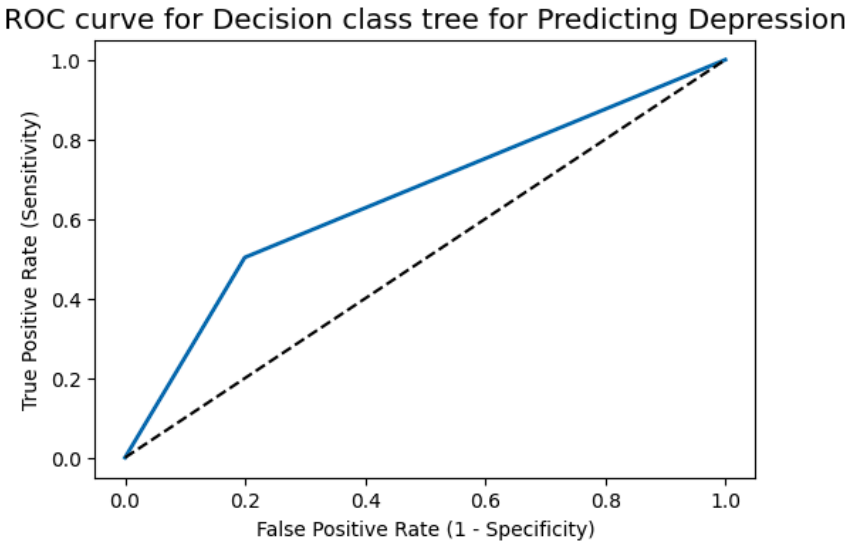

Figure 4.3.2.3: ROC curve of Decision tree classifier

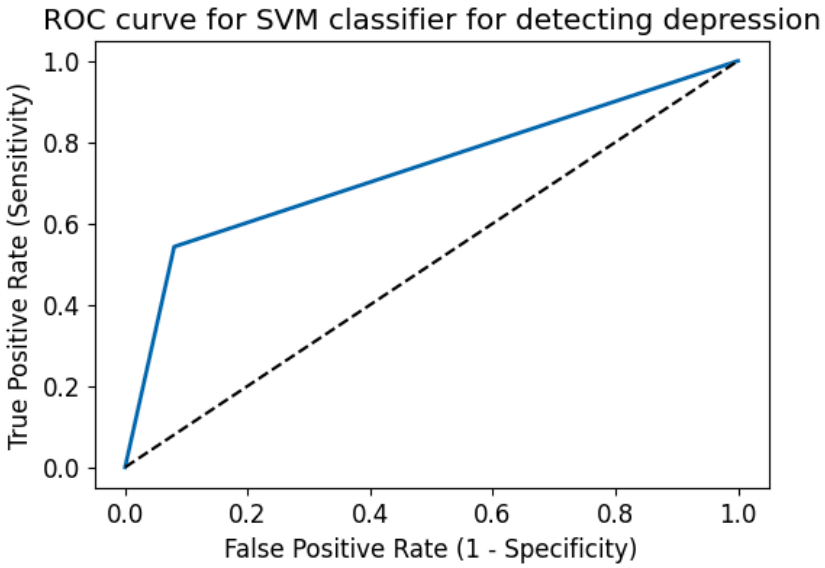

Figure 4.3.2.4: ROC curve of SVM

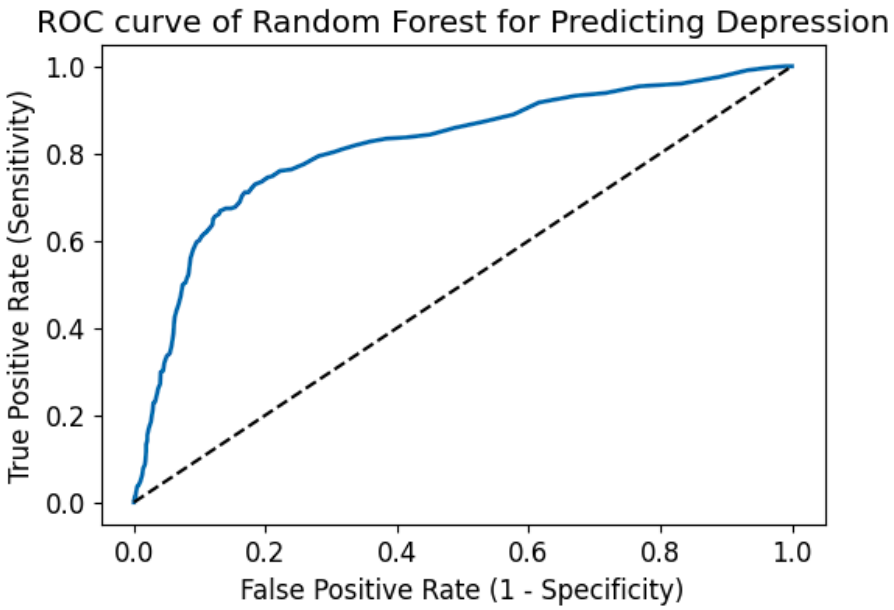

Figure 4.3.2.5: ROC curve of random forest

From all the ROC curves we can see that, Gaussian Naïve Bayes Classifier, KNN and random forest had similar performance.

4.3.3 Confusion Matrix

The confusion matrix is the way to assess the execution efficiency of a classification machine learning model.

The confusion matrix helps us visualize whether the model accurately distinguishes between two classes. As illustrated in the diagram below, it is a 2x2 matrix. The rows show the real values from the test set, while the columns represent the predicted values from the model.

For the Naïve Bayes model we checked to confusion matrix to see how well it did.

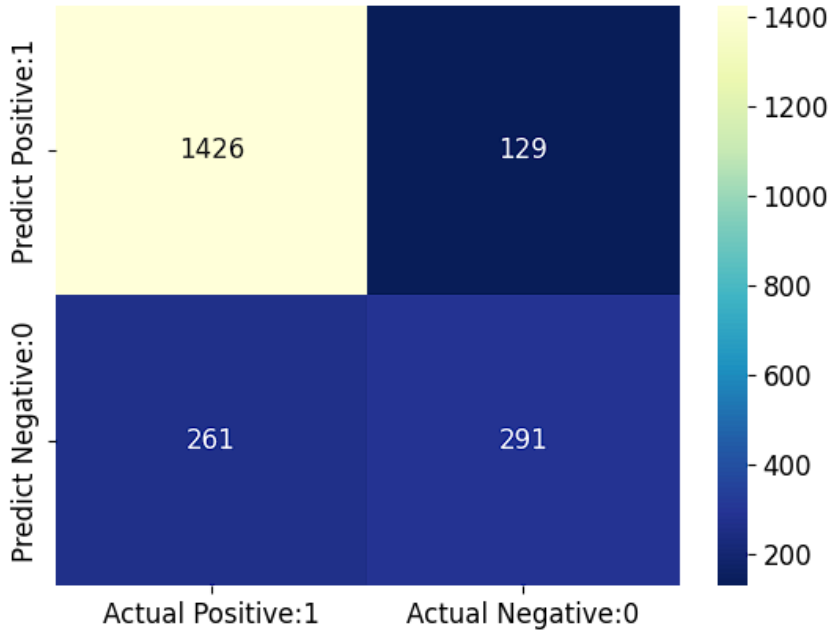

Figure 4.3.3: Confusion matrix of Naïve Bayes model

Top left element (True Positive): The model correctly classified Positive as Positive 1426 times.

Top right (False Negative): The model misclassified Negative as Positive 129 times.

Bottom left (False Positive): The model misclassified Positive as Negative 261 times.

Bottom right (True Negative): The model correctly classified Negative as Negative 291 times.

So, overall the model classified correctly 1717 times and misclassified 390 times. This is an overall good performance.

At the end, considering all things Naïve Bayes model has performed moderately with 83.72%. But, for clinical use the model needs to perform better.

4.3.4 Key Factors

By using the correlation function we can identify the key features in identifying the target feature of depression. We have found following features correlating the most with depression,

Stress_level 0.422782

Sleep_Quality 0.242975

Physical_Activity 0.015360

Diet_Quality 0.010067

These four key features helped the models more with detecting the target feature of depression. By employing the correlation function, we've conducted an in-depth analysis to identify the key features that have the most substantial impact on detecting depression as the target feature. Our findings have revealed that certain features exhibit stronger correlations with depression compared to others. Specifically, Stress_level emerges as the most influential feature, displaying a correlation coefficient of 0.422782. This suggests that higher levels of stress are associated with an increased likelihood of depression.

Additionally, Sleep_Quality demonstrates a moderate correlation coefficient of 0.242975, indicating that poor sleep quality may also contribute to the presence of depressive symptoms. On the other hand, features such as Physical_Activity and Diet_Quality exhibit weaker correlations with depression, with correlation coefficients of 0.015360 and 0.010067, respectively.

While these features still contribute to the overall predictive model, their influence on depression detection may be less pronounced compared to stress and sleep quality. Overall, these key features play a crucial role in enhancing the effectiveness of our models in identifying individuals at risk of depression, providing valuable insights for targeted intervention strategies and support programs.

CHAPTER 5

IMPLEMENTING IN UNIVERSITIES

5.1 Implementation

The Naïve Bayes model with an accuracy of 83.72% for the detection of depression and anxiety symptoms in university students can be considered quite promising, but its practical implementation in a university setting should be evaluated with caution. Here are some points to consider:

Strengths:

1. **High Accuracy**: An accuracy of 83.72% indicates that the model correctly identifies depression and anxiety symptoms in a significant majority of cases. This suggests that the model could be a useful tool for early detection and intervention.
2. **Potential for Early Intervention**: Implementing this model in a university setting could help in identifying students who may need mental health support, potentially leading to timely intervention and support.
3. **Resource Allocation**: The model can assist in prioritizing resources and efforts towards students who are most likely in need of assistance, thereby optimizing the use of mental health resources.

Considerations and Limitations:

1. **False Positives and False Negatives**: With an accuracy of 83.72%, there will still be a proportion of false positives (students incorrectly identified as having symptoms) and false negatives (students with symptoms not identified). It's important to evaluate the model's sensitivity (true positive rate) and specificity (true negative rate) to understand these implications better.
2. **Model Validation**: Before wide implementation, the model should be rigorously validated and tested in the specific university context. This includes checking for biases and ensuring it works well across diverse student populations.
3. **Complementary Use**: The model should not be used in isolation but rather as a supplementary tool to existing mental health services. It can flag potential cases, but

human judgment and professional mental health evaluations are essential for accurate diagnosis and treatment.

Practical Implementation:

1. **Pilot Testing**: Start with a pilot program to evaluate the model's effectiveness in the real-world university environment. Gather feedback from students and mental health professionals.
2. **Continuous Monitoring and Improvement**: Regularly monitor the model's performance and make necessary adjustments based on feedback and new data. Machine learning models can degrade over time if not properly maintained.
3. **Training and Awareness**: Train university staff on how to use the model and handle the data responsibly. Raise awareness among students about the tool and its purpose.

The implementation of a machine learning model to detect depression and anxiety symptoms in university students can have significant positive impacts, particularly in terms of early detection and resource optimization. However, careful consideration of privacy, ethical issues, and the potential for misidentification is essential to minimize negative impacts and ensure that the model effectively supports student mental health and well-being.

5.2 Ethical Aspects

Implementing a machine learning model for the detection of depression and anxiety symptoms in university students involves several ethical considerations that must be carefully addressed. Here are some key ethical aspects.

1. **Informed Consent**: Informed consent entails ensuring that students possess comprehensive knowledge about how their data will be utilized for mental health monitoring.

   For instance, before participating in a study utilizing their social media posts for sentiment analysis related to mental health, students should be provided with detailed information about the study, the types of data collected, and the potential risks involved. Explicit consent should be obtained from each student which will allow making informed decisions about their participation.

2. **Data Privacy and Confidentiality**: Protecting students' privacy involves implementing robust data protection measures and adhering to stringent confidentiality standards.

   For example, if medical records are used in a research to detect mental health symptoms, the data must be anonymised and encrypted to avoid illegal access. Access to the data should be restricted to authorized persons only, and tight processes should be implemented to guarantee compliance with any privacy requirements.

3. **Bias and Fairness**: Minimizing bias in machine learning models is essential for ensuring fairness and equity in their outcomes.

   For instance, when training a model to detect depression symptoms, it's crucial to use diverse and representative datasets that encompass a wide range of demographic characteristics. Continuous monitoring of the model's predictions across different demographic groups helps identify and address biases. Additionally, conducting fairness audits to evaluate the model's performance in diverse contexts helps ensure equitable outcomes.

4. **Transparency and Accountability**: Transparency in model deployment involves providing clear explanations of the model's algorithms, inputs, and decision-making processes.

   For example, if a machine learning model is used to predict mental health outcomes based on survey responses, students should be informed about the factors considered by the model and how their responses contribute to the predictions. Establishing clear accountability mechanisms, such as oversight committees or ethical review boards, helps ensure that ethical concerns are addressed promptly and effectively.

5. **Human Oversight and Intervention**: Human oversight and intervention play a crucial role in interpreting the outputs of machine learning models and providing appropriate support to students.

   For instance, if a model predicts a high likelihood of depression symptoms in a student based on their social media activity, trained mental health professionals should review the prediction and provide personalized interventions or referrals as necessary. Human

judgment ensures that ethical considerations are taken into account and that students receive appropriate support.

6. **Avoiding Stigmatization**: Minimizing the risk of stigmatization involves creating a supportive and non-judgmental environment for students seeking mental health support.

   For example, if a machine learning model is used to identify students at risk of depression based on academic performance data, it's essential to accompany the model's predictions with supportive interventions and resources. Emphasizing the confidentiality of mental health support services and reducing stigma through education and awareness campaigns fosters a culture of openness and acceptance.

7. **Empowerment and Autonomy**: Empowering students to make informed decisions about their mental health involves providing them with access to their own data and insights derived from machine learning models.

   For example, if a model predicts a student's likelihood of experiencing anxiety symptoms based on smartphone usage data, the student should have access to the prediction and the underlying data. Respecting students' autonomy and preferences in how they engage with mental health support services reinforces their agency and promotes self-advocacy.

8. **Continuous Monitoring and Evaluation**: Continuous monitoring and evaluation of machine learning models and their ethical implications are essential for ensuring ongoing compliance with ethical standards. For example, if a model is deployed to detect depression symptoms among students, regular evaluations should be conducted to assess its accuracy, fairness, and impact on student well-being. Soliciting feedback from students, mental health professionals, and other stakeholders helps identify areas for improvement and ensures that the model aligns with ethical principles over time.

By meticulously addressing these ethical aspects with detailed explanations and examples, university administrations can deploy machine learning models for mental health monitoring in a responsible and ethical manner, prioritizing student well-being while upholding principles of privacy, fairness, transparency, and accountability.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of the study

The study aims to develop and implement machine learning models for accurately detecting depression and anxiety symptoms in university students. Using data from surveys, medical records, social media, smartphones, and academic performance, the study seeks to identify key factors predictive of mental health issues through feature selection techniques, thereby improving model effectiveness and understanding.

Various machine learning algorithms including SVM, DT, RF, KNN, and NB are compared to determine the most reliable method for detecting depression and anxiety symptoms. Ethical concerns such as privacy and bias are addressed throughout the study, ensuring responsible and transparent practices in using machine learning for mental health assessment.

In conclusion, the study offers practical recommendations for implementing machine learning-based approaches in universities to support student well-being and academic success. These recommendations aim to enhance mental health support services while respecting student privacy and autonomy, ultimately contributing to a healthier and more supportive university environment.

6.2 Conclusion

In conclusion, the Naïve Bayes model demonstrated moderate performance with an accuracy of 83.72% in detecting depression and anxiety symptoms among university students. While this accuracy rate indicates promising potential for the model's application, several considerations must be taken into account regarding its implementation.

It is essential to recognize that the model's performance should be interpreted within the context of its intended use. While an accuracy of 83.72% is commendable, it is crucial to assess the model's sensitivity, specificity, and predictive values to understand its effectiveness comprehensively.

Additionally, the implementation of the Naïve Bayes model in university settings must be approached with careful consideration of ethical and practical implications. Privacy concerns regarding the collection and use of student data must be addressed transparently, ensuring that appropriate consent and data protection measures are in place.

Furthermore, it is imperative to recognize the limitations of the Naïve Bayes model and its potential impact on student well-being. While the model can serve as a valuable tool for early detection and intervention, it should not replace personalized, human-centered approaches to mental health support.

In summary, while the Naïve Bayes model has shown promising performance in detecting depression and anxiety symptoms among university students, its implementation requires careful consideration of ethical, privacy, and practical concerns.

6.3 Implication for further study

The implications for further study in this area are substantial and multifaceted. Here are some avenues for future research:

1. **Refinement of Machine Learning Models**: Further research could focus on refining existing machine learning models or developing new ones to enhance their accuracy and effectiveness in detecting depression and anxiety symptoms among university students. This could involve exploring advanced algorithms, incorporating additional data sources, or optimizing feature selection techniques. Refining a machine learning model is a comprehensive endeavor involving various strategies aimed at enhancing its performance and adaptability. It begins with meticulous feature engineering, where input data is scrutinized to select relevant features and engineer new ones to capture intricate patterns and relationships.

   Following this, hyperparameter tuning fine-tunes the model's configuration, adjusting parameters like learning rates and regularization strengths to strike a balance between bias and variance. Cross-validation techniques play a pivotal role in evaluating model performance, enabling iterative refinement of the model architecture and hyperparameters until satisfactory metrics are achieved. Additionally, leveraging ensemble methods can

enhance predictive accuracy and robustness by combining the collective wisdom of multiple models. Through this holistic approach encompassing feature engineering, hyperparameter tuning, cross-validation, and ensemble methods, machine learning models can be refined to achieve superior performance and adaptability across various real-world applications.

2. **Longitudinal Studies**: Conducting longitudinal studies to track students' mental health over time could provide valuable insights into the progression of depression and anxiety symptoms throughout their university experience. This could help identify risk factors, protective factors, and patterns of symptom development, informing more targeted intervention strategies. Rather than assessing the model's capabilities at a single point in time, longitudinal studying involves tracking its behavior and predictive accuracy across multiple iterations or periods.

   This approach enables data scientists to gain insights into how the model performs under varying conditions, data distributions, and environmental factors. By longitudinally studying the model, researchers can identify potential drifts or shifts in the data, detect patterns of model degradation or improvement, and implement timely adjustments or interventions to maintain optimal performance. Additionally, longitudinal studying facilitates the exploration of long-term trends and patterns in model behavior, providing valuable insights for ongoing refinement and optimization efforts. Overall, longitudinal studying plays a crucial role in ensuring the continued relevance, reliability, and effectiveness of machine learning models in real-world applications.

3. **Cultural and Contextual Factors**: Investigating how cultural and contextual factors influence the manifestation and detection of mental health issues among university students is essential for developing culturally sensitive and contextually relevant interventions. Research in this area could explore cross-cultural differences in symptom presentation, help-seeking behaviors, and the effectiveness of support services. This encompasses a broad spectrum of socio-cultural, environmental, and institutional factors that can significantly impact the development, deployment, and acceptance of such

models within diverse educational settings. Cultural factors may include cultural norms, beliefs, values, and attitudes towards mental health, as well as linguistic and socio-economic disparities that can affect access to mental health resources and willingness to seek help.

Contextual factors, on the other hand, encompass the institutional policies, academic pressures, social dynamics, and technological infrastructure present within university environments, all of which can influence the feasibility and appropriateness of implementing machine learning solutions for mental health assessment. By examining and addressing these cultural and contextual factors, the report aims to provide insights and recommendations for designing culturally sensitive and contextually relevant machine learning interventions that effectively support student well-being and academic success across diverse university settings.

4. **Integration of Technology:** With the increasing integration of technology into university settings, further study could explore the use of novel technologies, such as wearable devices, smartphone applications, or social media analytics, for mental health monitoring and intervention. This could open up new possibilities for early detection, personalized support, and remote intervention. This integration involves leveraging technological advancements to collect, analyze, and interpret data from various sources such as surveys, social media, smartphone applications, and academic records. By harnessing the power of machine learning, predictive models can be developed to identify patterns and indicators of mental health issues, facilitating early detection and intervention.

   Additionally, technology enables the implementation of scalable and accessible mental health support services, including online counseling platforms, digital self-help resources, and personalized interventions based on individual needs and preferences. However, successful integration of technology in mental health assessment requires careful consideration of ethical, cultural, and contextual factors to ensure inclusivity, privacy protection, and alignment with the diverse needs and values of university students. Through thoughtful integration of technology, this report aims to harness the potential of machine learning to enhance mental health support and promote student well-being in university settings.

5. **Effectiveness of Interventions:** Evaluating the effectiveness of interventions aimed at supporting student mental health is crucial for identifying best practices and improving outcomes. Future research could focus on assessing the impact of various interventions, such as counseling services, peer support programs, or resilience-building workshops, on student well-being and academic success.

   These interventions may include both preventive measures aimed at promoting mental well-being and targeted interventions designed to identify and support students experiencing depression and anxiety symptoms. Effectiveness is assessed based on a range of factors, including the accessibility, acceptability, and perceived usefulness of interventions by students, as well as their ability to improve mental health outcomes, reduce stigma, and enhance overall student success and academic performance. Interventions may encompass a diverse array of approaches, such as psych education, counseling services, peer support programs, mindfulness training, and technology-based interventions leveraging machine learning algorithms for early detection and personalized support.

   By evaluating the effectiveness of these interventions, the report aims to identify best practices and recommendations for developing evidence-based strategies that effectively support student mental health and well-being in university settings.

Overall, further study in these areas has the capability to advance our knowledge of student mental health, improve the effectiveness of support services, and ultimately contribute to creating healthier and more supportive university environments.

REFERENCE

[1]N. N. Moon, A. Mariam, S. Sharmin, M. M. Islam, F. N. Nur, and N. Debnath, "Machine learning approach to predict the depression in job sectors in Bangladesh," Current Research in Behavioral Sciences, vol. 2, p. 100058, Nov. 2021, doi: https://doi.org/10.1016/j.crbeha.2021.100058.

[2]J. Lin and Q. Chen, "ACADEMIC PRESSURE AND IMPACT ON STUDENTS' DEVELOPMENT IN CHINA," McGill Journal of Education, vol. 30, no. 002, pp. 5-10, 2019, doi: https://mje.mcgill.ca/article/view/8237

[3]M. S. H. Mukta, S. Islam, S. Shatabda, M. E. Ali, and A. Zaman, "Predicting academic performance: Analysis of students' mental health condition from social media interactions," Behavioral Sciences, vol. 12, no. 4, p. 87, Apr. 2022, doi: https://doi.org/10.3390/bs12040087.

[4]E. Steigerwald et al., "Overcoming Language Barriers in Academia: Machine Translation Tools and a Vision for a Multilingual Future," BioScience, vol. 72, no. 10, pp. 988–998, Aug. 2022, doi: https://doi.org/10.1093/biosci/biac062.

[5]M. Gil, S.-S. Kim, and E. J. Min, "Machine learning models for predicting risk of depression in Korean college students: Identifying family and individual factors," *Frontiers in Public Health*, vol. 10, Nov. 2022, doi: https://doi.org/10.3389/fpubh.2022.1023010.

[6] S. Ware et al., "Predicting depressive symptoms using smartphone data," Smart Health, vol. 15, p. 100093, Mar. 2020, doi: https://doi.org/10.1016/j.smhl.2019.100093.

[7] Laura et al., "Predicting stress in first-year college students using sleep data from wearable devices," PLOS digital health, vol. 3, no. 4, pp. e0000473–e0000473, Apr. 2024, doi: https://doi.org/10.1371/journal.pdig.0000473.

[8] S. Saeb, E. G. Lattie, S. M. Schueller, K. P. Kording, and D. C. Mohr, "The relationship between mobile phone location sensor data and depressive symptom severity," PeerJ, vol. 4, p. e2537, Sep. 2016, doi: https://doi.org/10.7717/peerj.2537.

[9] C. L. Bae, D. Mills, F. Zhang, M. Sealy, L. Cabrera, and M. Sea, "A Systematic Review of Science Discourse in K–12 Urban Classrooms in the United States: Accounting for Individual, Collective, and Contextual Factors," Review of Educational Research, vol. 91, no. 6, pp. 831–877, Sep. 2021, doi: https://doi.org/10.3102/00346543211042415.

[10] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, “A survey on addressing high-class imbalance in big data,” Journal of Big Data, vol. 5, no. 1, Nov. 2018, doi: https://doi.org/10.1186/s40537-018-0151-6.

[11] C. Xu, P. Coen-Pirani, and X. Jiang, “Empirical Study of Overfitting in Deep Learning for Predicting Breast Cancer Metastasis,” Cancers, vol. 15, no. 7, pp. 1969–1969, Mar. 2023, doi: https://doi.org/10.3390/cancers15071969.

[12]P. Probst, A.-L. Boulesteix, and B. Bischl, “Tunability: Importance of Hyperparameters of Machine Learning Algorithms,” Journal of Machine Learning Research, vol. 20, no. 53, pp. 1–32, 2019, doi: https://www.jmlr.org/papers/v20/18-444.html

[13]A. Nassar and M. Kamal, “Ethical Dilemmas in AI-Powered Decision-Making: A Deep Dive into Big Data-Driven Ethical Considerations,” International Journal of Responsible Artificial Intelligence, vol. 11, no. 8, pp. 1–11, Aug. 2021, doi: https://neuralslate.com/index.php/Journal-of-Responsible-AI/article/view/43

[14] F. -L. Fan, J. Xiong, M. Li and G. Wang, "On Interpretability of Artificial Neural Networks: A Survey," in IEEE Transactions on Radiation and Plasma Medical Sciences, vol. 5, no. 6, pp. 741-760, Nov. 2021, doi: 10.1109/TRPMS.2021.3066428.

[15] S. Awadalla, E. B. Davies, and C. Glazebrook, “A longitudinal cohort study to explore the relationship between depression, anxiety and academic performance among Emirati university students,” BMC Psychiatry, vol. 20, no. 1, Sep. 2020, doi: https://doi.org/10.1186/s12888-020-02854-z.

[16]“Beck Depression Inventory | Online Psychological Inventories,” terappin.com. https://terappin.com/en/test/beck-depression-inventory

[17]“What is Ordinal Data? Definition, Examples, Variables & Analysis | Simplilearn,” Simplilearn.com, Sep. 29, 2022. https://www.simplilearn.com/what-is-ordinal-data-article

[18]W. F. Scientist Data, “One-Hot Encoding — A Brief Explanation,” Medium, Aug. 03, 2023. https://medium.com/@WojtekFulmyk/one-hot-encoding-a-brief-explanation-8c5daec395e3

[19]Z. Ali and Sb. Bhaskar, “Basic Statistical Tools in Research and Data Analysis,” Indian Journal of Anaesthesia, vol. 60, no. 9, p. 662, Sep. 2016, doi: https://doi.org/10.4103/0019-5049.190623.

[20]Y.-Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," Shanghai archives of psychiatry, vol. 27, no. 2, pp. 130–5, 2015, doi: https://doi.org/10.11919/j.issn.1002-0829.215044.

[21]S. Dash, "Decision Trees Explained — Entropy, Information Gain, Gini Index, CCP Pruning..," Medium, Nov. 02, 2022. https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c

[22]Z. Zhang, "Introduction to machine learning: k-nearest neighbors," Annals of Translational Medicine, vol. 4, no. 11, pp. 218–218, Jun. 2016, doi: https://doi.org/10.21037/atm.2016.03.37.

[23]"k-Nearest Neighbors Algorithm - an overview | ScienceDirect Topics," www.sciencedirect.com. https://www.sciencedirect.com/topics/computer-science/k-nearest-neighbors-algorithm

[24]IBM, "What is the k-nearest neighbors algorithm? | IBM," www.ibm.com, 2023. https://www.ibm.com/topics/knn

[25]S. Huang, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," Cancer Genomics & Proteomics, vol. 15, no. 1, Jan. 2018, doi: https://doi.org/10.21873/cgp.20063.

[26]S. Winters-Hilt, A. Yelundur, C. McChesney, and M. Landry, "Support Vector Machine Implementations for Classification & Clustering," BMC Bioinformatics, vol. 7, no. S2, Sep. 2006, doi: https://doi.org/10.1186/1471-2105-7-s2-s4.

[27]Shuzhan Fan, "Understanding the mathematics behind Support Vector Machines," Shuzhan Fan, May 07, 2018. https://shuzhanfan.github.io/2018/05/understanding-mathematics-behind-support-vector-machines/

[28]C. Savas and F. Dovis, "The Impact of Different Kernel Functions on the Performance of Scintillation Detection Based on Support Vector Machines," Sensors, vol. 19, no. 23, p. 5219, Nov. 2019, doi: https://doi.org/10.3390/s19235219.

[29]S. Abdullah and H. F. Eid, "Utilizing random forest algorithm for early detection of academic underperformance in open learning environments," PeerJ, vol. 9, pp. e1708–e1708, Nov. 2023, doi: https://doi.org/10.7717/peerj-cs.1708.

[30]J. Wolfson et al., “A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data,” Statistics in Medicine, vol. 34, no. 21, pp. 2941–2957, May 2015, doi: https://doi.org/10.1002/sim.6526.

[31]M. Ontivero-Ortega, A. Lage-Castellanos, G. Valente, R. Goebel, and M. Valdes-Sosa, “Fast Gaussian Naïve Bayes for searchlight classification analysis,” NeuroImage, vol. 163, pp. 471–479, Dec. 2017, doi: https://doi.org/10.1016/j.neuroimage.2017.09.001.

Plagiarism Report

DETECTION OF DEPRESSION AND ANXIETY SYMPTOMS

ORIGINALITY REPORT

20%	17%	10%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Submitted to Daffodil International University Student Paper	2%
3	export.arxiv.org Internet Source	<1%
4	www.hindawi.com Internet Source	<1%
5	fastercapital.com Internet Source	<1%
6	arxiv.org Internet Source	<1%
7	Submitted to University of Sheffield Student Paper	<1%
8	www.researchsquare.com Internet Source	<1%
9	journals.plos.org Internet Source	<1%