Ideation of Depression and Suicide Using Machine Learning Techniques

Final Year Design Project

By
Student Name: Aspy Rihan
Student ID: 211-15-3979

Final Year Design Project Report
This Report Presented in Partial
Fulfillment of the Requirement for the Degree of
Bachelor of Science in Computer Science & Engineering

Supervised by

Ms. Taslima Ferdaus Shuva
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by

Ms. Hasnur Jahan
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Daffodil International University
Dhaka, Bangladesh

14 May, 2025

APPROVAL

This Project titled **"Ideation of Depression and Suicide Using Machine Learning Technology"**, submitted by Aspy Rihan ID No: **211-15-3979** 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 B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **14 May, 2025**.

BOARD OF EXAMINERS

Ms. Nazmun Nessa Moon (NNM) — **Chairman**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Shah Md Tanvir Siddiquee (SMTS) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

14.05.2025

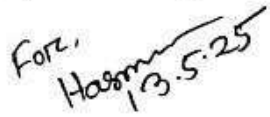
Mr. Md Umaid Hasan (MUH) — **Internal Examiner**
Sr. Lecture
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Ahmed Wasif Reza (DWR) — **External Examiner**
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

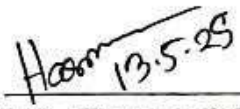
We hereby declare that this project has been done by us under the supervision of **Ms. Taslima Ferdaus Shuva Assistant Professor** Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Ms. Taslima Ferdaus Shuva
Assistant Professor
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by:



Ms. Hasnur Jahan
Lecturer
Department of Computer Science and EngineeringDaffodil International University

Submitted by:

Aspy Rihan
Student ID:211-15-3979
Department of Computer Science and EngineeringDaffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Ms. Taslima Ferdaus Shuva Assistant Professor** Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Machine Learning** carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

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

ABSTRACT

Depression has emerged as a major mental health challenge globally, with a noticeable rise in prevalence in Bangladesh, particularly accompanied by increasing suicidal tendencies. This study investigates the underlying causes of depression and presents a machine learning-based approach for its early detection. Unemployment, family pressure, work stress, and social isolation were identified as key contributing factors.

This issue was addressed using several supervised machine learning models, including Nave Bayes, Support Vector Machine (SVM), Logistic Regression (LR), Random Forest, Linear Discriminant Analysis (LDA), AdaBoost, Decision Tree, and k-Nearest Neighbor (k-NN). A comprehensive dataset related to mental health and depression symptoms was used to train and test these models. Statistical Machine Learning has shown to be the most accurate and consistent method while Logistic Regression provided the most consistent and balanced performance. Naïve Bayes had good recall capabilities, and AdaBoost had robust performance across a variety of metrics.

Additionally, Random Forest and k-NN provided reliable results, while Decision Tree and LDA did not produce any interpretable yet effective results. This study confirms the potential of machine learning techniques for the accurate detection of depression and related mental health issues. It will be important to enhance model explanation, reduce algorithmic bias, integrate diverse data sources, and adhere to ethical principles like privacy protection and informed consent for future research. In resource-constrained regions like Bangladesh, AI-driven mental health tools are especially important for enabling timely diagnosis and support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Depression, which is one of the most common mood disorders, negatively affects the mental state of millions of people across the nation, manifesting itself by persistently depressed mood, disinterest, and the inability to carry through any activities. The term depression comes from the Latin word “deprimere” [1], which means to press down, and it has been written about as far back as ancient scriptures from Hippocrates, and remains one of the biggest medical challenges.

According to the World Health Organization, more than 720,000 people die by suicide every year. A report compiled by the World Health Organization (WHO) showed that more than 264 million people [2], or 3.4 percent of the world's population, suffer from depression, with 15-29-year-olds being the most affected. The increasing observed trend has also exacerbated the occurrence of high suicide rates, where 800,000 people [3] take their own lives every year making suicide the fourth main cause of death among the youth. It is estimated that 20 people attempt suicide for every suicide that occurs. Statistics such as these are not just numbers. They represent countless tears, unspoken pain, and shattered dreams.

Suicide, however, can be prevented. We can address this crisis by raising awareness, encouraging open discussion, and ensuring proper mental health support. The state, families, and society can work together to create a compassionate environment where those in distress are no longer alone.

Using clinical, behavioral, and demographic data, machine learning (ML) and data analytics have opened new possibilities for identifying high-risk individuals. In this study, the primary objective is to analyze. The use of machine learning techniques to predict suicide risk based on depression-related data.

1.2 Motivation

There has been a significant increase in depression and suicidal tendencies in recent years, especially among the younger generation. Bangladesh National Institute of Mental Health and the World Health Organization (WHO) conducted a survey in Dhaka revealing that over 18% of children and adolescents suffer from depression. The WHO anticipates that depression may create a global socio-economic crisis by 2030. Researchers believe that one in every five individuals may experience depression at some point in their life.

Depression often leads to suicide. Suicide is the act of ending one's own life. Many countries consider suicide a crime, and a person who attempts suicide can be sent to jail. According to the WHO, 20% of those who commit suicide do so using toxic substances or firearms. In Bangladesh, 20,505 people died by suicide in 2023, with an average of 56 deaths per day. In the previous year, 2016, the suicide rate was 23,868, with 14.7 suicides per 100,000 people.

There are various reasons why an individual may decide to take their own life, including mental health issues like depression, financial crises, family disputes, and the loss of loved ones. Several risk factors are associated with suicide, such as:

- Lack of access to mental health services
- Victimization by sexual abuse
- Family history of suicide
- Mental health issues like depression or personality disorders
- Excessive use of alcohol or drugs

Addressing and preventing such issues requires the use of modern technology. Machine learning has emerged as a powerful tool in this regard. Machine learning technologies can analyze textual, behavioral, and demographic data to identify individuals at risk. Through Natural Language Processing (NLP), sentiment analysis, and other classification algorithms, machine learning models can detect subtle indicators in text, such as loneliness, depression, or loss of interest in life.

If this initiative is implemented, it will not only benefit society but also enhance my technical skills and knowledge. As a technology student, gaining experience in applying technology to solve real-world problems is extremely valuable. Through this, I will not only gain expertise in machine learning and data analysis but also develop as a socially conscious technologist with a humanistic perspective.

1.3 Research Questions

- Which age group is most affected by depression and why?
- Under what circumstances do individuals tend to consider suicide?
- How do factors like age, gender, and occupation influence suicidal tendencies?
- How effective is social media data in predicting the risk of depression or suicide?
- How do the volume and diversity of data affect model accuracy and generalization?
- Which machine learning algorithms are most effective for detecting depression, and why?

1.4 Objectives

This research seeks to accomplish several important goals in identifying suicidal tendencies among adolescents through the application of machine learning techniques.

1. Early Identification and Timely Support:

The primary aim of this research is to develop a reliable machine learning model capable of detecting individuals at risk of suicide in its early stages. Early identification allows mental health professionals to intervene promptly with appropriate support, significantly reducing the likelihood of suicidal actions.

2. Assessment of Model Effectiveness:

To evaluate how accurately the proposed models can predict suicidal tendencies in the future and assess their overall performance in real-world scenarios.

3. Integration of Diverse Data Sources:

To explore the impact of combining multiple data types—such as textual, behavioral, and demographic data—into machine learning models for enhancing the accuracy and reliability of depression and suicide risk predictions.

4. Regional Analysis and Model Suitability:

To analyze and determine which prediction models perform best at forecasting suicide rates across different continents or regions, considering geographical and cultural variations.

5. Identification of the Most Efficient Predictive Model:

With the help of existing datasets and relevant evaluation criteria, this research aims to assess and determine the most reliable and high-performing predictive model for anticipating suicide rates.

1.5 Project Outcome

The following outcomes are expected after the successful completion of this study:

1. **Development of a Predicative Model**

   Identifying individuals with depression or suicidal tendencies In order to improve accuracy, we will develop a robust machine learning model. Combination This model will be trained and tested using text, behavioral, and demographic data.

2. **Data-Driven Insights into Mental Health**

   This project will examine a variety of datasets, including survey data, and demographic information, in order to uncover key patterns and risk factors associated with depression and suicide. Healthcare professionals and policy makers can use these insights to make informed decisions.

3. **Evaluation of Algorithm Performance**

   The research will provide a comparative analysis of several machine learning algorithms (e.g., Decision Trees, SVM, Random Forest, Naive Bayes, Deep Learning models) to identify which techniques are most effective for mental health

prediction, along with their limitations.

4. **Prototype Implementation**

   A working prototype or application may be created to demonstrate how such a predictive system can be deployed in real-world scenarios—such as monitoring mental health trends on social media or within specific communities.

5. **Contribution to Mental Health Awareness and Prevention**

   By highlighting the potential of technology in early identification and prevention of mental health crises, this project will contribute toward building a more responsive and compassionate mental health support system.

6. **Scope for Future Enhancement**

   The findings and model design will create a foundation for future research and improvement, such as the inclusion of real-time monitoring, voice/speech analysis, or integration with mental health support platforms and mobile applications.

1.6 Organization of the Report

This thesis is organized into the following chapters, each building on the previous one to provide a comprehensive view of the research, methodology, and outcomes:

- **Chapter 1: Introduction**

  This chapter introduces the context of the research, highlighting the motivation behind studying depression and suicidal tendencies through machine learning. It outlines the research questions, objectives, and scope of the study, as well as providing a roadmap of the thesis structure to guide the reader through the upcoming chapters.

- **Chapter 2: Background Study**

  Chapter 2 presents a comprehensive examination of prior research and scholarly contributions related to mental health and the application of machine learning. It explores earlier developments in predictive systems for mental health, outlines the primary obstacles encountered in this domain, and emphasizes the existing research gaps—especially concerning the use of machine learning to forecast depression and

suicidal tendencies. The chapter concludes with a synthesis of the reviewed literature, thereby providing a strong rationale for the chosen research methodology.

- **Chapter 3: Methodology**

  Chapter 3 outlines the structured methodological framework followed throughout this research. It encompasses detailed descriptions of the data gathering procedures, the preprocessing steps undertaken to clean and normalize the data, and the specific feature extraction techniques applied to enhance model performance. Various machine learning algorithms are explored and justified based on their relevance and efficiency in the context of mental health prediction.

- **Chapter 4: Implementation, and Results**

  A detailed description of software, hardware, and the step-by-step process of developing the predictive model is presented in Chapter 4. The paper explains how the algorithms were run, including preprocessing and feature engineering. The chapter compares model performances using standard metrics and highlights key insights that show the system's potential for real-world application. Overall, the results support the research goals and contribute to mental health prediction through machine learning.

- **Chapter 5: Engineering Standards and Design Challenges**

  The engineering principles and challenges faced during the project are discussed in Chapter 5. It ensures the system is compliant with software, hardware, and communication standards. Using machine learning for mental health also raises ethical concerns, societal impacts, and sustainability concerns. As a final step, it describes the engineering solutions applied and future plans for scaling the system.

- **Chapter 6: Conclusion**

A summary of the research's main findings and contributions is presented in the final chapter. There is a discussion of the study's limitations and some suggestions for future improvement. The paper concludes with recommendations for further research and advances in the field, including ways to expand and apply the work.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

The purpose of this section is to provide a comprehensive review of existing research on mental health and machine learning. The study examines the goals, methods, and findings of past studies rather than just summarizing them. Additionally, it examines the challenges those researchers faced and how they impacted their results. By showing how past efforts have shaped current understanding and pointing out areas that need further study, it helps build a strong foundation for our own work. In addition to reviewing the literature, we also discuss the specific challenges we encountered during our research. There were technical issues, limited or poor-quality data, time constraints, and unexpected issues during implementation. We explain what happened and how we adjusted our approach for each challenge.

An honest reflection provides a clear picture of the research process, allowing readers to appreciate the complexities involved.Furthermore, it offers useful insights and strategies for future researchers who may encounter similar obstacles, contributing to ongoing discussions on mental health prediction and machine learning.

2.2 Literature Review

Authors	Year	Method/Approach	Data Source	Key Findings	Focus Area
Mulay et al.	2020	Visual-based depression detection using facial features	Visual input	Effective non-invasive depression detection via facial expression analysis	Depression detection
Ding et al.	2020	Deep Integrated SVM (DISVM) combining deep learning with SVM	Sina Weibo posts	Improved precision and generalizability of hybrid models	Depression detection
Birjali et al.	2017	Hybrid of ML and semantic sentiment analysis	Social media content	Predicted suicide-related sentiments effectively; need for contextual understanding	Suicide risk detection
Bernert et al.	2020	Review of AI/ML in suicide prevention	Multiple sources	Highlighted ethical, clinical, and technical challenges in real-world use	Suicide prevention
Tadesse et al.	2019	Deep learning for suicide ideation detection	Social media forums	Neural networks effective in real-time suicide risk identification	Suicide detection
Wilimitis et al.	2022	Integration of clinical screening with ML	Clinical and real-time data	Improved accuracy of early intervention using combined methods	Suicide risk prediction
Valdez et al.	2020	ML for suicide detection in Spanish social networks	Spanish language social media	Language and cultural factors affect model performance	Suicide detection (Spanish context)
Yeskuatov et al.	2022	ML & NLP with Reddit data	Reddit	Platform-specific data improves suicide ideation detection	Suicide detection on Reddit
Shukla et al.	2020	Speech-based depression detection	Audio signals	Voice features are viable indicators for mental health	Depression detection via speech
Patel et al.	2019	CBT-based intelligent chatbot	Student interactions	Chatbots can support emotional well-being and provide mental health assistance	AI therapy/chatbots
Dahiwade et al.	2019	ML model for disease prediction	Medical datasets	Data-driven models aid in early diagnosis and medical intervention	General disease prediction

Mbarek et al.	2019	Twitter-based suicidal profile detection	Twitter posts	Behavioral & linguistic analysis identifies at-risk individuals	Suicide detection
Rahman et al.	2020	ML for depression detection	Bangladeshi university students	SVM showed highest accuracy; psychosocial features are key	Depression among students
Gil et al.	2022	ML for depression prediction	Korean college students	Random Forest performed best; family and self-perception factors were significant	Depression prediction
Cruz et al. (1st mention)	2023	Naïve Bayes classifier	519 university students	78.03% accuracy; good sensitivity for moderate/severe cases	Student depression detection
Roy Chowdhury et al.	2025	I-HOPE (interpretable ML)	Five-year longitudinal student data	91% accuracy; emphasizes personalized assessments	Depression prediction in students
Lopes and Nihei	2024	Systematic review of ML for mental health	48 studies on students	ML models are promising for remote/underserved mental health care	Student mental health detection
Cruz et al. (2nd mention)	2023	Naïve Bayes (repeat entry)	519 university students	Same as above	Same as above
Bhaumik and Stange	2023	RE-EM trees & MERF for vulnerability detection	Young adult psychological factors	Identified subgroups at greatest risk using complex psychological feature interactions	Personalized depression risk modeling

Table 2.1: Literature Review

2.3 Comparative analysis

To understand the landscape of depression and suicide detection research, a comparative analysis of several existing studies was conducted. These studies vary widely in terms of methodology, data sources, and application domains. While some works focus on text-based analysis from social media platforms (e.g., Reddit, Twitter), others leverage audio, video, or chatbot interactions to detect mental health conditions. Machine learning and deep learning remain common approaches across all studies, with an increasing trend toward multimodal data analysis.

Table 2.2: Comparative analysis

Authors (Year)	Methodology	Description	Result
Yeskuatov et al. (2022)	ML + NLP on Reddit data	Reviewed Reddit posts to explore suicidal ideation patterns	Highlighted importance of social media in early detection
Mbarek et al. (2019)	Linguistic and behavioral analysis on Twitter	Analyzed tweets to identify suicidal tendencies based	Demonstrated linguistic cues are effective for sulcidal risk
Mulay et al. (2020)	Computer vision-based emotion detection	Used facial expressions from visual inputs to detect depression	Showed visual signs can accurately indicate depresion
Shukla et al. (2020)	Energy & statistical feature extraction from speech	Analyzed speech energy levels and statistics to identify depression	Validated speech features as indicators of mental health
Patel et al. (2019)	Chatbot using Cognitive Behavioral Therapy	Developed an AI chatbot to interact with students and assess depression	Showed ML can predict diseases, including mental health issues

The table below summarizes the methodologies, descriptions, and results of key papers, highlighting their strengths and the gaps that remain in the literature—particularly the lack of Bengali-language focused datasets and systems, which this research aims to address.

2.4 Gap Analysis

Sl. No.	Identified Gap	Description	How This Research Addresses It
1	Limited Regional Data Representation	Most existing datasets are from Western countries and do not reflect local socio-cultural contexts like Bangladesh.	Uses region-specific data or localized case studies to improve relevance and accuracy in South Asian contexts.
2	Lack of Social Media Data Integration	Social media platforms contain real-time emotional cues often ignored in traditional research.	Incorporates social media data (text/posts) using NLP and sentiment analysis to detect signs of depression early.

3	Use of Single Data Type	Many studies use only one data type (text, behavior, or demographics), limiting model accuracy.	Combines textual, behavioral, and demographic data to create a more robust and holistic predictive model.
4	Generalized Models for All Age Groups	Current models often do not focus on the most vulnerable age group—adolescents and young adults.	Focuses specifically on youth (15–29 years), who are statistically more prone to depression and suicidal tendencies.
5	Black-Box Nature of Models	Lack of explainability makes it difficult for mental health professionals to trust and use the results.	Explores interpretable models and feature importance to make predictions understandable and actionable.
6	Ignorance of Region-Specific Risk Factors	Factors like academic stress, family pressure, and financial instability are often ignored.	Includes region-specific risk variables in the data to reflect the real-life context in developing countries.

Table 2.3: Grap Analyses

2.5 Scope of the Problem

Our research work primarily involves analyzing the given data and applying machine learning algorithms to create a model. In our society, many people face mental health issues, but they do not realize that they are depressed. They feel isolated and do not know what to do. In such cases, they need a system that can help identify their problem and assist them in understanding whether they are truly depressed. This system will provide them with appropriate guidance so that they can take the right steps and restore their mental well-being.

2.6 Summary

This chapter has laid the foundation for understanding the research problem, its significance, and the justification for conducting this study. It began with an overview of the rising global and regional concern surrounding depression and suicide, particularly among youth. The

An emphasis is placed on the urgent need for technological solutions in the motivation section, which explains both personal and societal reasons for choosing this topic.The motivation section explains both personal and societal reasons for choosing this topic, emphasizing the urgent need for technological solutions. The research questions were carefully designed to explore how demographic factors relate to suicidal tendencies and to evaluate how well machine learning can predict these risks. The objectives set the study's scope and expected outcomes, including early detection, comparing algorithms, and analyzing regional differences. A comparative analysis reviewed various approaches, tools, and models previously used to detect depression, pointing out their strengths and weaknesses. This gap analysis reveals key limitations in current research, such as a lack of region-specific data, limited use of real-time social media inputs, and difficulties explaining how predictive models make decisions. These insights highlight the importance and originality of this study. In summary, this chapter places the research within the wider body of knowledge, clearly defining its scope, significance, and expected contributions. It lays the groundwork for the following chapters, which will cover research methodology, data collection, model development, and experimental analysis.

CHAPTER 3

Research Methodology

3.1 Methodology/Requirement Analysis & Design Specific cation

This research aims to develop a predictive model that can detect signs of depression and assess suicidal tendencies based on daily activities and behavioral responses. Methodology, requirements, and design specifications are outlined in this section, which ensures a systematic and clear approach to achieving established goals.

3.1.1 Methodology Overview

Human behaviors, emotions, and motivations are analyzed using machine learning (ML). As well as linguistic cues derived from the collected dataset. Several components are included in the methodology

In stages:

- Data Collection: Collecting a comprehensive dataset that includes behavioral responses and personal information pertinent to depression and suicide risk.
- Data Preprocessing: Cleaning, refining, and transforming raw data into a structured format suitable for machine learning applications.
- Model Selection and Training: Implementing various machine learning algorithms to develop predictive models.
- Model Evaluation: Evaluating the performance of each model using a range of metrics to gauge accuracy and reliability.
- Interpretation: Analyzing the results and ensuring that the outputs of the model are interpretable and meaningful.

3.1.2 Requirement Analysis

In the requirement analysis phase, both functional and non-functional requirements are identified to ensure that the developed model meets both technical specifications and user requirements.

A. Functional Requirements

- **Data Input**: The system must accept structured data inputs such as age, gender, profession, marital status, and behavioral responses to predefined questions
- **Prediction**: The system should classify individuals into two categories: depressed and non-depressed.
- **Machine Learning Algorithms**: Implement multiple algorithms to compare and evaluate depression risk prediction, including:
  - k-Nearest Neighbor (kNN)
  - Logistic Regression
  - Support Vector Classifier (SVC – Linear)
  - Naïve Bayes
  - Random Forest
  - Adaptive Boosting (AdaBoost)
  - Decision Tree
  - Linear Discriminant Analysis (LDA)
- **Evaluation Metrics**: Models should be evaluated for accuracy, precision, recall, and F1-value Score, and AUC-ROC
- **Output**: In order to determine depression and suicidal tendencies, the model should provide clear, interpretable results.

B. Non-Functional Requirements

- **Scalability**: The system must be scalable to handle large datasets for future implementations.
- **Data Security**: It is essential to ensure that personal data is securely protected, maintaining confidentiality and safeguarding privacy at all times

- **User Experience**: The system interface must be designed to be user-friendly and intuitive, enabling mental health professionals to interact with it effortlessly and efficiently.
- **Efficiency**: The system should provide predictions in real-time or near-real-time with minimal delay.
- **Performance**: The system must provide highly accurate predictions based on the data provided.

3.1.3 Data Collection and Preprocessing

In order to analyze depression and suicidal tendencies, a Google Form was used to collect personal information and behavioral patterns. The data collection process was carefully planned to ensure reliability; however, privacy concerns and participant cooperation emerged as challenges.

Data Features

A total of 28 features are included in the dataset, including:

1. Age
2. Gender
3. Profession
4. Marital status
5. Behavioral responses to depression and emotional well-being queries..

Preprocessing Steps

- **Data Cleaning**: Handling missing or null values in the dataset.
- **Normalization**: Normalizing text data and converting categorical variables into numerical formats.
- **Feature Selection**: Identifying key features by evaluating their importance using relevant metrics.
- **Data Splitting**: Dividing the dataset into training, testing, and validation subsets to guarantee an unbiased assessment of the model's performance.

Serial No	Processing Step	Description
1	Raw Dataset	Collected Raw Data
2	Checking Null Values	Identify Missing (null)Values
3	Handing Null Values	Impute or Remove Missing Values
4	Label Encoding	Convert Categorical Data to Numeric
5	Normalization	Scale Data to a Standard Range
6	Feature Engineering	Create or Select Meaningful Features
7	Drop Outcome	Remove Target Variable for Independent Prep
8	Processed Dataset	Final Cleaned and Transformed Dataset

Table 3.1: Data processing Table

3.1.4 Machine Learning Model Design

In order to ensure the robustness and accuracy of depression and suicidal tendencies predictions, several machine learning models were selected. These models were selected for their ability to efficiently handle classification tasks and uncover patterns within the dataset. The models include:

1. **k-Nearest Neighbor (kNN)**: A non-parametric algorithm used to classify data points based on their proximity
2. **Logistic Regression**: A linear classifier for binary classification.
3. **Support Vector Classifier (SVC – Linear)**: is a powerful statistical machine learning tool that maximizes the margin between classes.
4. **Naïve Bayes**: A probabilistic classifier that performs well with text-based data.
5. **Random Forest**: An ensemble model that optimizes accuracy by combining predictions from multiple decision trees

6. **Adaptive Boosting (AdaBoost)**: A boosting technique that improves performance by combining weak learners
7. **Decision Tree**: A decision tree is a tree-based classifier that analyzes data based on feature values.
8. **Linear Discriminant Analysis (LDA)**: A method used in both dimensions reduction and classification tasks.

In order to determine which model performs best in identifying depression and suicidal tendencies, several metrics will be evaluated, including accuracy, precision, recall, F1-score, and AUC-ROC.

3.1.5 Tools and Technologies

Several platforms and software tools were used to implement and analyze the predictive models, selected for their compatibility with machine learning algorithms and their ability to handle large datasets efficiently:

- **Google Colab**: Cloud-based platform for running computationally demanding models and handling vast datasets, with free GPU access.
- **Jupyter Notebook**: An open-source environment for executing code, conducting data analysis, and visualizing results
- **Microsoft Excel**: Initially used for creating the dataset and analyzing preliminary data
- **Weka Software**: A graphical tool for visualizing data and machine learning models, supporting both classification and regression tasks.

3.1.6 Model Evaluation

To assess the reliability and effectiveness of the models, the following evaluation metrics will be utilized:

- **Accuracy**: The proportion of correctly classified instances in the dataset.
- **Precision**: The ratio of true positive predictions to all positive predictions.
- **Recall**: The ratio of true positive predictions to all actual positives.

- **F1-Score**: The harmonic mean of precision and recall, providing a balanced metric for evaluation.
- **AUC-ROC**: A performance measure for classification tasks at various threshold settings, evaluating the model's ability to discriminate between classes.

This structured methodology ensures a comprehensive and rigorous process for developing a predictive model that has significant potential to contribute to mental health research, especially by identifying individuals at risk of suicide and depression. Each model will undergo cross-validation and comparison to determine the most accurate and reliable approach to predicting depression and suicidal tendencies. With multiple machine learning algorithms, thorough evaluation, and advanced tools, the research offers valuable insights for early intervention.

3.1.7 Proposed Methodology

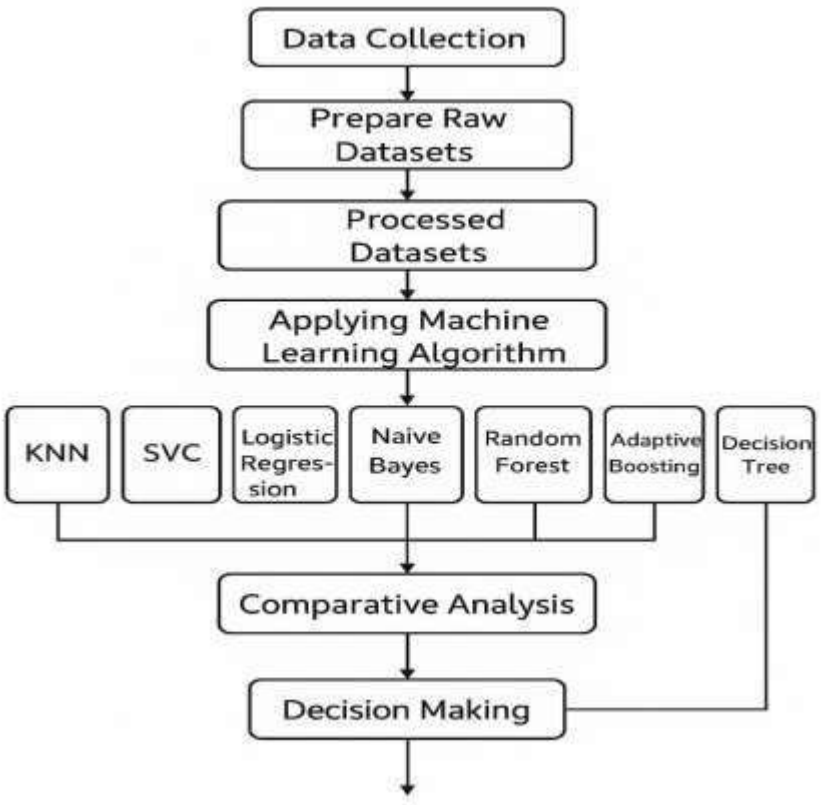


Fig 3.2: Proposed Methodology

3.1.8 Functional and Non-functional Requirements

Functional Requirements:

1. **Data Input and Collection**:

   The platform should be capable of processing structured inputs, such as demographic data and behavioral responses. Information should be collected through digital forms or online surveys.

2. **Prediction and Classification**:

   Analyzing the provided input data, the system should be capable of identifying signs of depression and suicidal ideation. It should deliver a binary classification result, indicating whether or not the User is depressed.

3. **Machine Learning Algorithms**:

   For prediction, the system should utilize k-NN, Logistic Regression, Naive Bayes, and Random Forest models Each algorithm will generate a prediction, and the system will select the most accurate model.

4. **Model Training and Evaluation**:

   The system should train and evaluate each model based on several metrics such as accuracy, precision, recall, and F1-score to ensure reliable performance. Cross-validation should be implemented to make the models robust and generalizable.

5. **Data Storage and Retrieval**:

   All input data should be stored along with corresponding predictions for future analysis, model improvement, and auditability.

6. **Reporting and Visualization**:

   In order to summarize prediction accuracy and model performance metrics, the system should display visual results, such as charts and graphs.

Non-functional Requirements:

1. **Scalability**:
    - The system should be scalable to handle large datasets and a growing number of users.
2. **Security**:
    - Ensure the protection of sensitive personal information and the use of secure communication channels for data collection.
3. **Performance**:
    - The system should provide real-time predictions with minimal latency.
4. **Usability**:
    - The interface should be user-friendly, with clear instructions for inputting data and interpreting results.
5. **Reliability**:
    - The system should have minimal downtime, with high availability and data integrity.
6. **Maintainability**:
    - The code and infrastructure should be easy to maintain, with clear documentation.

3.1.9 Context Diagram

A **Context Diagram** is a high-level visual representation of the system and its interactions with external entities (users, systems, etc.). The key components in the diagram are:

- **External Entities**:
    - **User (Mental Health Professional)**: Inputs data and reviews predictions.
    - **Google Form or Data Collection Platform**: Collects input data such as demographics and behavioral responses.
- **System**:
    - **Depression Prediction System**: Receives input, processes it with ML algorithms, and generates predictions.

- **Output**:
    - The system outputs a **prediction** (depressed / not depressed) and **visualization of results**.

This diagram can be created using diagram tools like Lucid chart, Microsoft Visio, or any similar tool.

3.2 Quantitative Data Analysis

As part of our study, we collected data from a total of 508 individuals belonging to various classes, professions, and social backgrounds. The participants represented a diverse range of age groups, marital statuses (married and unmarried), and both male and female genders. Based on this demographic information, we created a pie chart, which is presented in Figure 3.3. This chart visually illustrates the distribution of depression among the respondents. According to the data, approximately 80.4% of the participants are experiencing some form of depression, while only 19.6% are not currently facing depressive symptoms. This significant disparity highlights the widespread prevalence of mental health issues—particularly depression—within the sampled population.

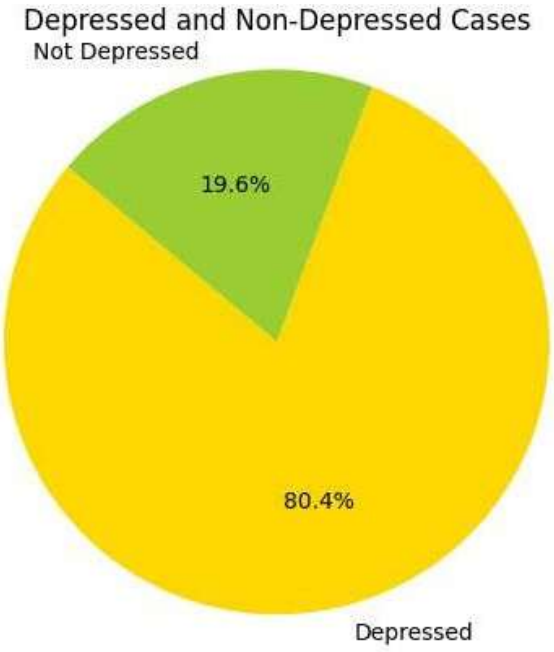


Fig 3.3: Depressed and non-depressed

`

In the next step of our analysis, we explored suicidal tendencies among the respondents. Each individual was asked whether they had ever attempted suicide. Based on their responses, we generated another pie chart, shown in Figure 3.4. This chart reveals that 52.2% of the participants reported having never attempted suicide, while a concerning 47.8% admitted to having made at least one suicide attempt. These findings are alarming, as they indicate the severity of depression's impact and how it can lead individuals to consider or even act on suicidal thoughts.

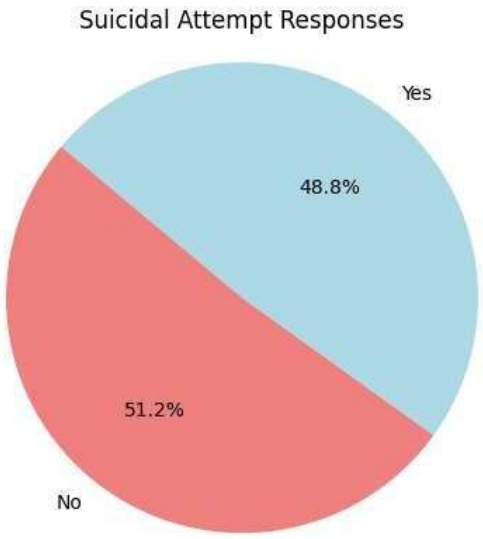


Fig 3.4: Suicidal Attempt Responses

These two visual analyses serve as crucial foundations for our research. They help us to better understand the depth and consequences of mental health conditions like depression and provide insight into the level of suicidal behavior linked with such conditions. This information is not only vital for the development of our model but also for identifying the root causes of the problem and designing preventive interventions in the future.

The analysis of Figure 3.5 reveals that the rate of depression is higher among females compared to males. According to the data, 205 males are experiencing depression, while 40 males are not. In contrast, 202 females are suffering from depression, and 59 females are not. This clearly indicates that females are more affected by depression than males. The disparity may be attributed to various social, familial, and cultural pressures that women often face, which can lead to increased psychological stress. These findings highlight the need for greater awareness and support for women's mental health in society.

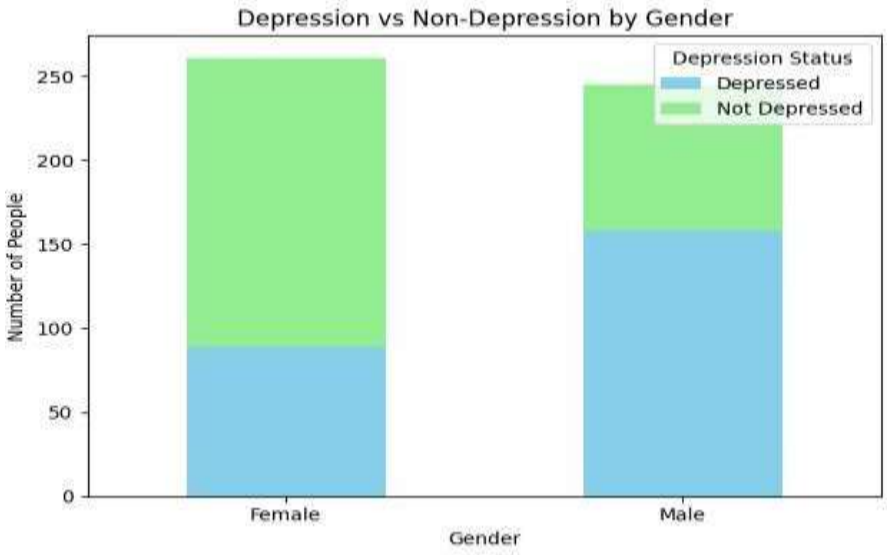


Fig 3.5: Depression Vs Not-Depression by Gender

According to Figure 3.6, a significant gender-based difference is observed in the prevalence of suicidal thoughts. Among 261 female participants, 89 reported having suicidal thoughts, while the remaining 172 did not. In contrast, out of 245 male participants, 158 expressed experiencing such thoughts, whereas 87 did not. This data indicates that suicidal ideation is more prevalent among males, despite the slightly higher number of female respondents.

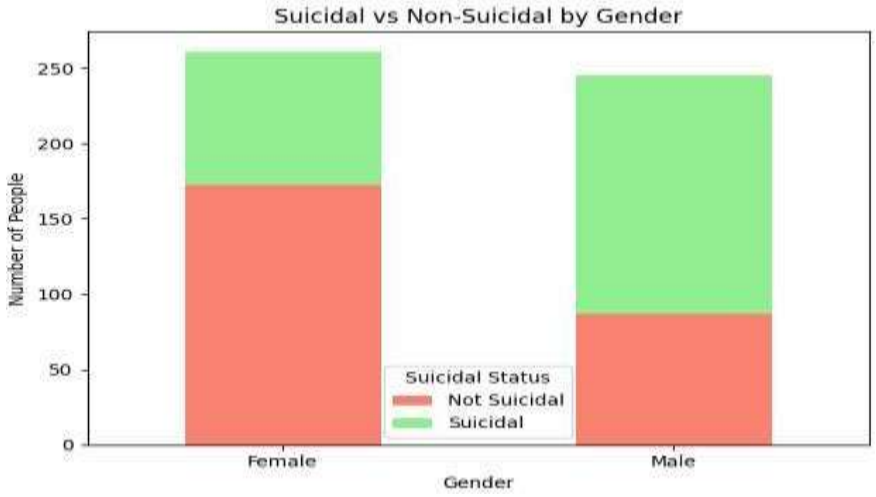


Fig 3.6: Suicidal Vs Not-Suicidal by Gender

Analysis of Figure 3.7 shows that individuals aged between 21 and 25 are the most affected by depression. Compared to other age groups, this age range has the highest number of people experiencing depressive symptoms. This phase of life often involves significant stress due to academic pressure, career uncertainty, identity formation, and social or familial expectations. These factors can contribute to a heightened vulnerability to mental health issues, making young adults in this age group more susceptible to depression.

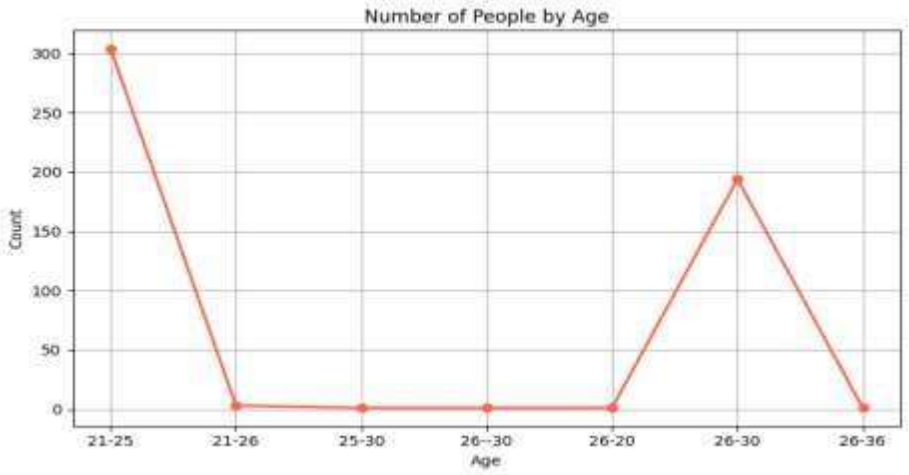


Fig 3.7: Number of People by Age

Figure 3.8 illustrates which professions have the highest rates of depression. It clearly shows that individuals in certain professions—such as students or private sector employees—are more likely to experience depression compared to others. On the other hand, Figure 3.9 presents data on which professions have the highest rates of suicidal tendencies. It reveals that professions associated with higher levels of stress tend to show a greater likelihood of suicide attempts.

Figure 3.8 highlights a clear variation in depression levels across different occupational groups. Among the employed participants, 62 individuals were found to be experiencing symptoms of depression. In the case of the unemployed, the number rises to 159. However, the most affected group appears to be students, with 245 reporting signs of depression.

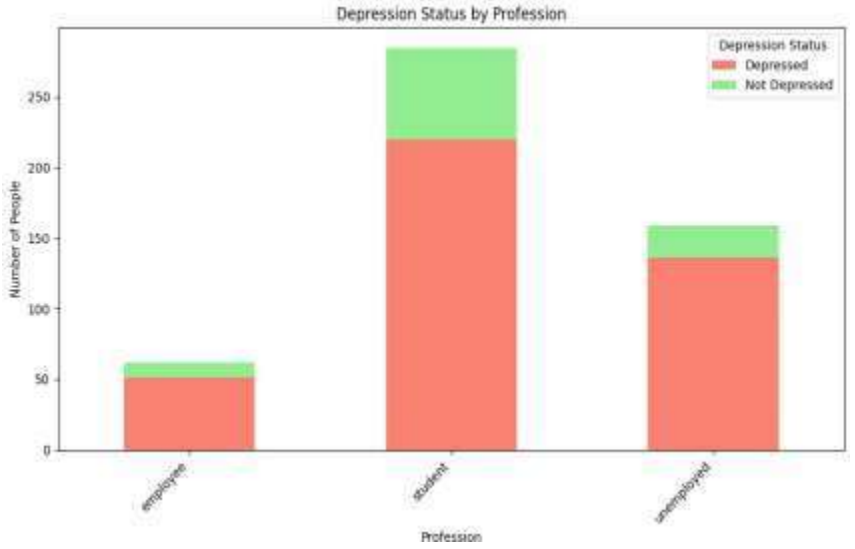


Fig 3.8: Depression Status by Profession

According to Figure 3.9, the suicide rates vary across different occupational groups. Among employed individuals, 22 cases of suicide were recorded. In contrast, the number is significantly higher among the unemployed, with 82 cases. The highest number of suicides was reported among students, totaling 143 cases. These figures highlight the varying levels of psychological and social pressure experienced by different groups in society.

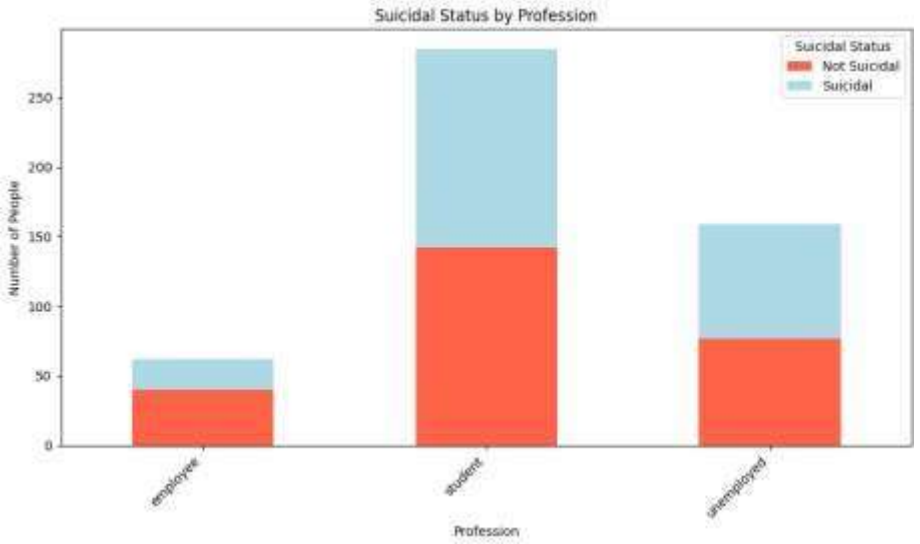


Fig 3.9: Suicide Status by Profession

Collectively, these two visual representations offer a comprehensive understanding of the disparities in mental health challenges among various occupational sectors.

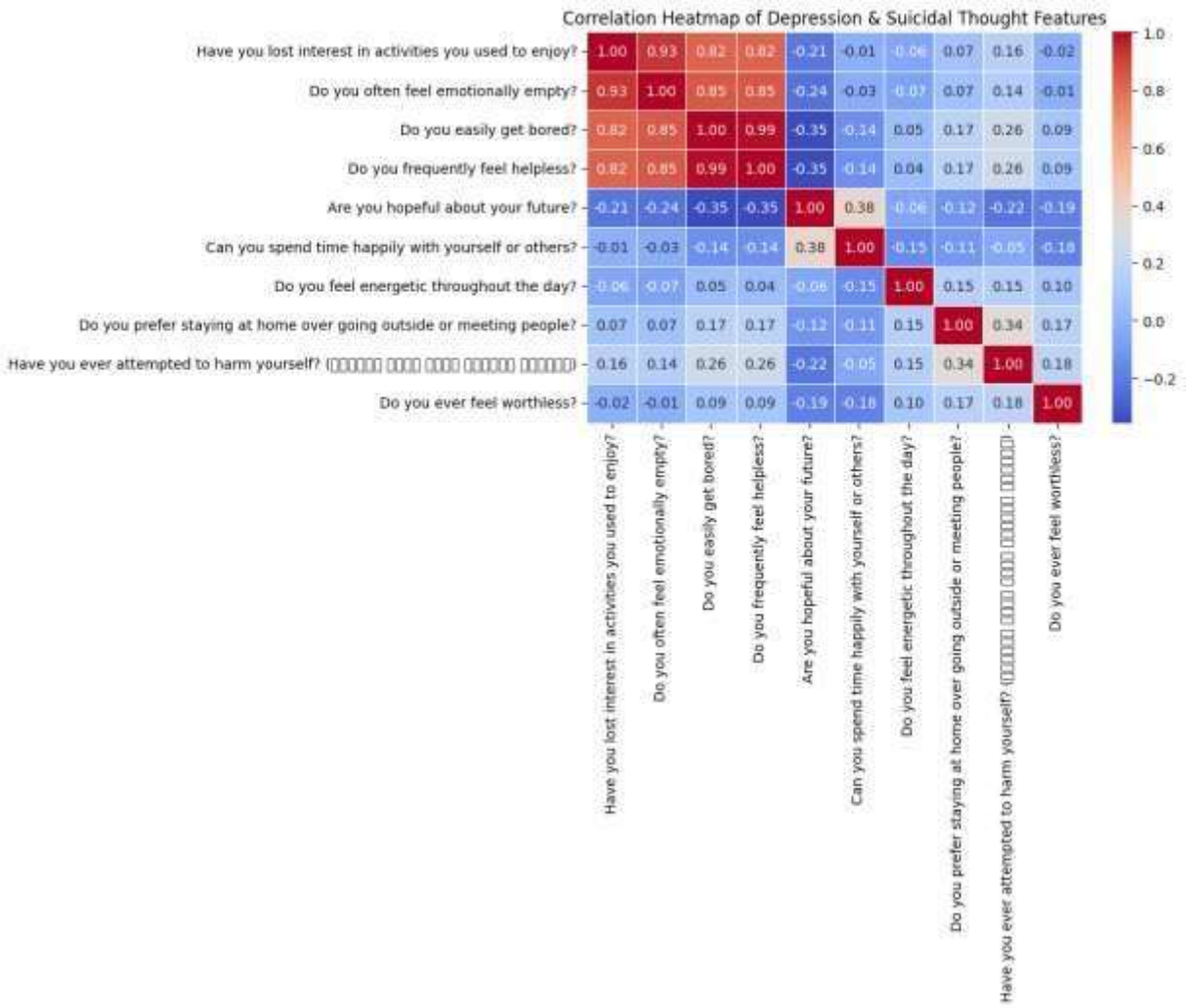


Fig 3.10: Correlation Matrix

The correlation heatmap above presents the relationships among various features related to depression and suicidal ideation. Analyzing the chart reveals that several psychological indicators are closely linked with each other, reflecting recognizable patterns commonly associated with depressive states.

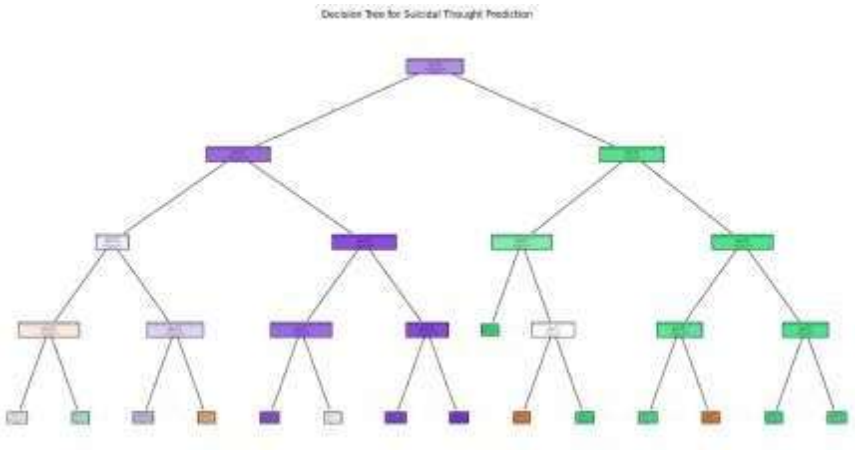


Fig 3.10: Decision Tree

The Decision Tree model illustrated in Figure 3.10 helps identify individuals at risk of suicide. Each leaf node displays the distribution of classes along with the number of corresponding samples. Among the 15 leaf nodes in the tree, several are associated with a predicted class of '1', which typically indicates a potential risk of suicidal thoughts. By summing the sample counts from those specific leaf nodes, we can estimate the total number of individuals who are likely at risk. This aggregation serves as a crucial part of the model's analytical insight.

3.3 Project Plan

The project has been structured into distinct phases to ensure systematic execution and adherence to the proposed timeline.

Stage 1: Data Collection (1 Month)

- Collect data through surveys.
- Clean and preprocess the data.

Stage 2: Model Development (2 Months)

- Implement machine learning algorithms.
- Train and evaluate models using the dataset.

Stage 3: Model Optimization and Evaluation (1 Month)

- Optimize the models for performance.
- Evaluate model accuracy, precision, recall, etc.

Stage 4: Final Implementation and Testing (1 Month)

- Integrate models into a final system.
- Conduct system testing with users.

Stage 5: Report and Documentation (1 Month)

- Finalize the thesis.
- Prepare the report and presentation.

3.4 Project Outcome

The findings from this study are expected to offer meaningful insights into the mental health conditions prevalent within distinct demographic categories. By pinpointing which segments of the population are more prone to depression or carry a heightened risk of suicidal behavior, the research aims to contribute to the development of more focused and impactful mental health strategies.

The anticipated contributions of this study include:

1. **Identification of High-Risk Groups**: Recognizing specific age brackets, professional backgrounds, or social circumstances that are disproportionately affected by mental health issues.
2. **Data-Driven Policy Development**: Equipping healthcare professionals and policymakers with empirical evidence to better allocate mental health resources and implement targeted outreach programs.
3. **Community Awareness and Early Intervention**: Promoting mental health literacy

within communities to foster timely intervention, reduce stigma, and encourage open conversations.

4. **Enhanced Support Services**: Informing the creation of specialized mental health programs and counseling services designed to address the needs of the most vulnerable populations.
5. **Foundation for Future Inquiry**: Establishing a base for subsequent research aimed at exploring more complex psychological, social, or behavioral factors associated with mental health.

In essence, this research seeks not only to deliver statistical knowledge but also to contribute to practical, real-world improvements in how mental health is understood, managed, and supported.

3.5 Summary

This chapter outlines the methodological framework, system requirements, and architectural design principles employed throughout the research. The study utilizes a machine learning–driven approach to detect depressive symptoms and potential suicidal tendencies. It includes an in-depth exploration of both functional and non-functional requirements, the system's data flow, and the overall project development lifecycle. Various machine learning models are implemented and evaluated to identify the most effective predictive algorithm. Ultimately, this methodological approach is intended to facilitate early mental health assessment and contribute to the development of proactive intervention tools.

Chapter 4

Implementation and Results

4.1 Introduction

Multiple machine learning classifiers were evaluated on the curated dataset in detail in this section. We considered the following algorithms in this study: Support Vector Classifiers (SVC), Random Forests, K-Nearest Neighbors (kNN), Linear Classifiers, Nave Bayes, Adaptive Boosting, Logistic Regression, Decision Trees, and Linear Discriminant Analysis (LDA). We identified the model whose prediction accuracy was the highest.

A two-phase evaluation strategy was used to accomplish this. Initially, each algorithm was tested on the raw dataset without any preprocessing or feature enhancement to establish baseline performance. In the following step, the dataset was preprocessed and features were engineered, and the models were reevaluated. To identify the most effective predictive approach and assess the impact of feature engineering on accuracy, this comparative analysis was conducted.

4.2 Environment Setup

A proximity-based classification algorithm leveraging k-Nearest Neighbors (kNN) was also explored, but its computational complexity increases with data volume. As a linear classifier, the Stochastic Gradient Descent Classifier (SGD) was evaluated for its efficiency in processing large datasets and its effectiveness in linearly separable problems. It was used in scenarios that involved smaller or text-based datasets due to its simplicity and speed. To further improve model performance, AdaBoost (Adaptive Boosting) was applied to iteratively improve weak learners by focusing on previously misclassified data. Because of the probability-based output and interpretability of logistic regression, a standard binary classification tool was chosen. With hierarchical data splitting, the Decision Tree algorithm provided a transparent and easily interpretable method for making predictions. Lastly, Linear Discriminant Analysis (LDA) was incorporated into the analysis in order to maximize class separation under linear assumptions by projecting data. A set of performance indicators was used to evaluate each model, including Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and ROC Curve. By comparing these

metrics, we were able to identify the most effective model for predicting depression.

4.3 Testing and Evaluation/Performance/ Comparative Analyses

A dataset containing 24 selected behavioral, emotional, and demographic features was used to evaluate the predictive models.To determine the effectiveness of machine learning algorithms in predicting depression or not depression, several algorithms were trained and tested. As a result of the evaluation, most models achieved high accuracy levels, demonstrating the reliability of the dataset and feature selection strategy. As a result of its consistent efficiency in detecting depression-related patterns, the Support Vector Classifier (SVC) achieved an accuracy rate of approximately 92%. The k-Nearest Neighbor (kNN) and Linear Discriminant Analysis (LDA) models demonstrated superior performance, achieving accuracy rates of 94.82% and 96.49%, respectively. As a result, they are able to analyze and replicate the underlying structure of the data more effectively.Despite its minimalistic design, Na'ive Bayes accomplished an accurate accuracy of 92.94%, highlighting its effectiveness despite its somewhat simple probabilistic framework.

indicates their enhanced ability to analyze and replicate the underlying structure of the data.

Meanwhile, the Naïve Bayes classifier, known for its relatively simple probabilistic framework, delivered a dependable accuracy of 92.94%, highlighting its effectiveness despite its minimalistic design.

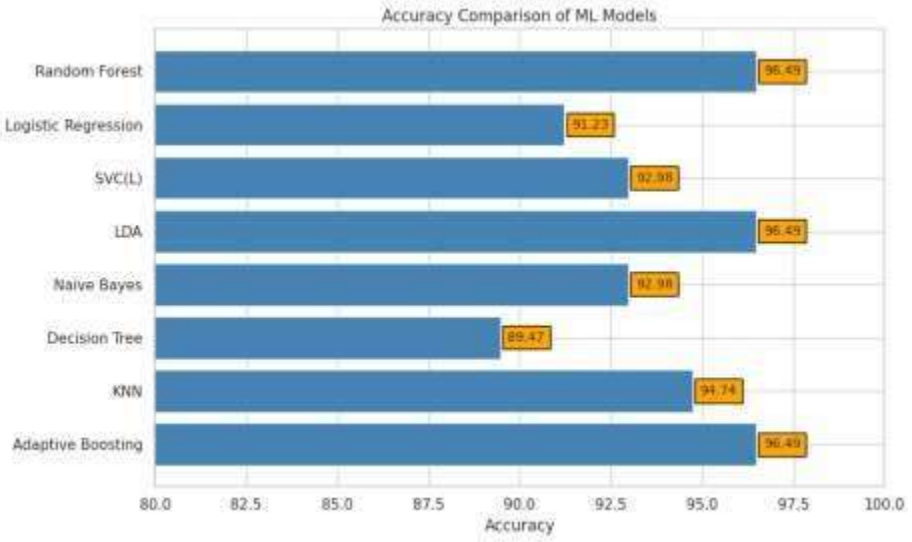


Fig 4.1: Accuracy of Model Predicting Depression

According to the analysis presented in Figure 4.2, the Random Forest model demonstrated the highest level of performance, achieving an accuracy rate of 94.74%. The linear discriminant analysis (LDA) and adaptive boosting models both achieved an accuracy rate of 92.98%. As a result of the Logistic Regression and Naive Bayes models, each achieved 91.23% accuracy, whereas the Decision Tree and K-Nearest Neighbor (kNN) models were each 89.47% accurate. Based on these results, the Random Forest model emerged as the most effective among all evaluated classifiers, with an accuracy of only 84.21%. The lowest performance was observed in the SVC (Linear) model, with an accuracy of only 84.21%.

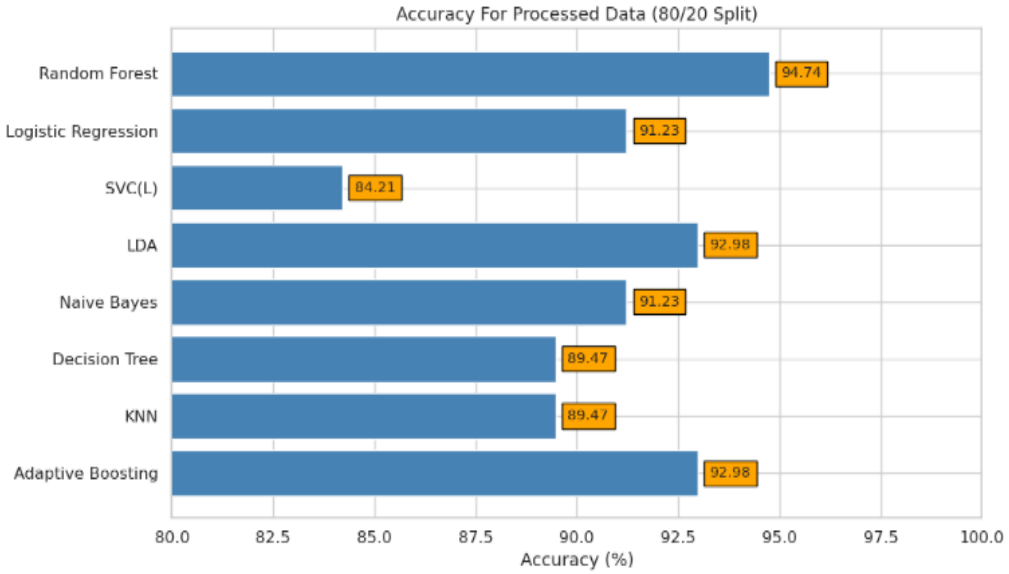


Fig 4.2: Accuracy of Model Predicting Depression 80/20

As shown in Figure 4.3, where the dataset was split into 70% for training and 30% for testing, models like Random Forest, Logistic Regression, LDA, Decision Tree, KNN, and Adaptive Boosting achieved 100% accuracy on the test data. This indicates that these models were highly effective in capturing the underlying patterns in the data. On the other hand, the SVC (with a linear kernel) and Naive Bayes models showed slightly lower accuracy, both achieving 97.78%. While this is still a high level of performance, it is marginally lower compared to the other models. The slight drop in performance could be due to certain assumptions made by these algorithms. For instance, Naive Bayes assumes independence between features, which may not hold true in this dataset. Similarly, SVC may struggle if the data is not perfectly linearly separable, impacting its performance.

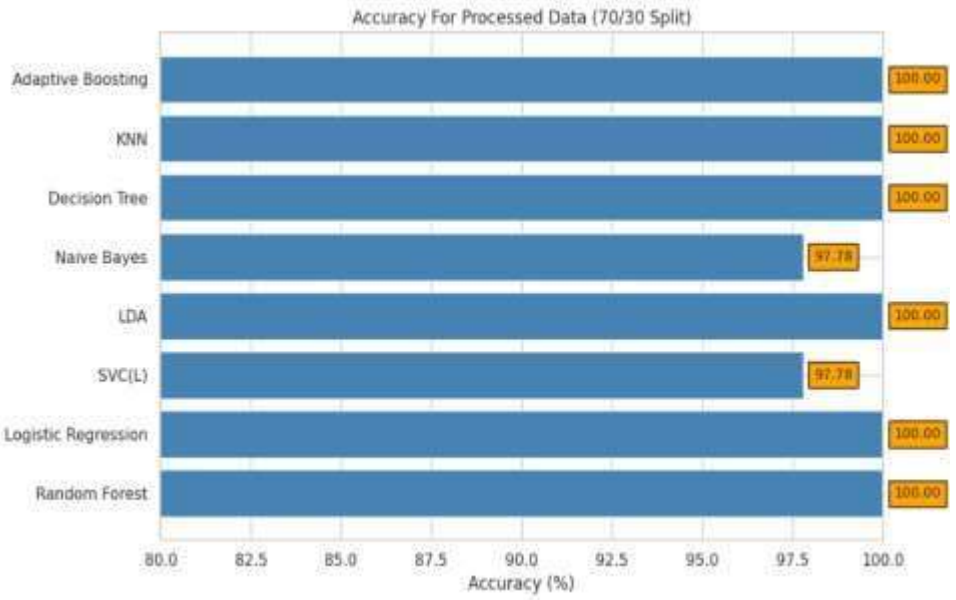


Fig 4.3: Accuracy of Model Predicting Depression 70/30

The KNN and LDA models achieved the highest accuracy, both reaching 96.55%. These models were most effective in capturing the patterns in the processed data. The Random Forest model also performed well with an accuracy of 93.10%, showing its strength as an ensemble method that combines multiple decision trees. Adaptive Boosting, Naive Bayes, and Linear SVC all yielded the same accuracy of 89.66%. These models provided moderate performance, possibly due to limitations in handling the data's complexity or feature interactions. On the other hand, Decision Tree and Logistic Regression showed the lowest accuracy at 86.21%. This suggests that these simpler models were less suited to the structure and distribution of the processed data used in this evaluation.

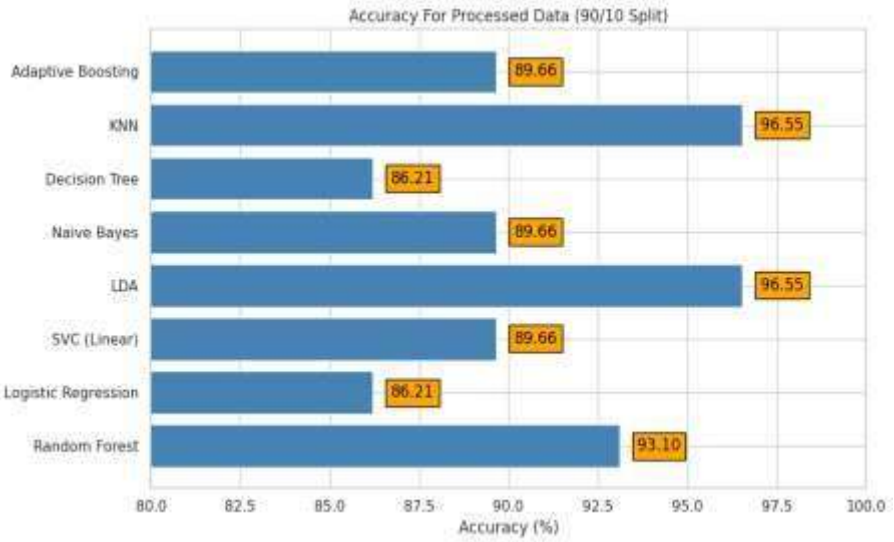


Fig 4.4: Accuracy of Model Predicting Depression 90/10

Decision Trees are extensively discussed in Section 18.3. The book explains how the algorithm builds a tree by recursively splitting data based on attribute values to reduce entropy (using information gain). Algorithms like ID3 and C4.5 are mentioned, along with techniques like pruning to avoid overfitting. The section also touches on the computational aspects and tree optimization [26]. Naive Bayes appears under Probabilistic Learning. It is based on Bayes' Theorem and assumes independence between features (the "naive" assumption). Despite its simplicity, it often performs surprisingly well, especially in domains like text classification. The book covers its probabilistic formulation, assumptions, and limitations [26]. LDA is not explicitly detailed in these pages, but it is conceptually related to topics covered under Pattern Recognition and Classification. LDA projects data in a way that maximizes class separation, based on statistics like class means and within-class variance. While the book focuses more on Bayesian and probabilistic methods, the mathematical background of LDA (e.g., covariance, linear separability) aligns with the discussions [26]. Support Vector Machines are introduced as part of Large Margin Classification. The book discusses how SVMs create hyperplanes with maximum margin between classes. While it mostly explains the conceptual basis and introduces kernel methods, Linear SVC is a specific, simplified version of this idea. Soft margin classifiers and

their optimization are also explored [26]. Logistic Regression is covered under Probabilistic Learning as a linear classifier that uses the logistic function to model the probability of a binary outcome. It explains how logistic regression fits data by maximizing the likelihood and includes discussion on gradient descent, overfitting, and regularization techniques [26]. Random Forests are not discussed by name in the book, but their foundation is. It is an ensemble method based on Bagging (Bootstrap Aggregating), where multiple decision trees are trained on random subsets of the data. The book discusses how ensemble methods can reduce variance and improve generalization—key ideas behind Random Forests [26].

Model	Accuracy (90/10 Split)	Accuracy (80/20 Split)	Accuracy (70/30 Split)	Overall Accuracy
Random Forest	93.10%	94.74%	100.00%	96.49%
Logistic Regression	86.21%	91.23%	100.00%	91.23%
SVC (Linear)	89.66%	84.21%	97.78%	92.98%
LDA	96.55%	92.98%	100.00%	96.49%
Naive Bayes	89.66%	91.23%	97.78%	92.98%
Decision Tree	86.21%	89.47%	100.00%	89.47%
KNN	96.55%	89.47%	100.00%	94.74%
Adaptive Boosting	89.66%	92.98%	100.00%	96.49%

Table 4.5: Accuracy Table

Several curves corresponding to different algorithms are shown in the figure. The figure illustrates the ROC curves for multiple classification algorithms. Each curve represents the performance of a specific model, with its corresponding AUC (Area Under the Curve) value displayed. A higher AUC indicates better model performance.

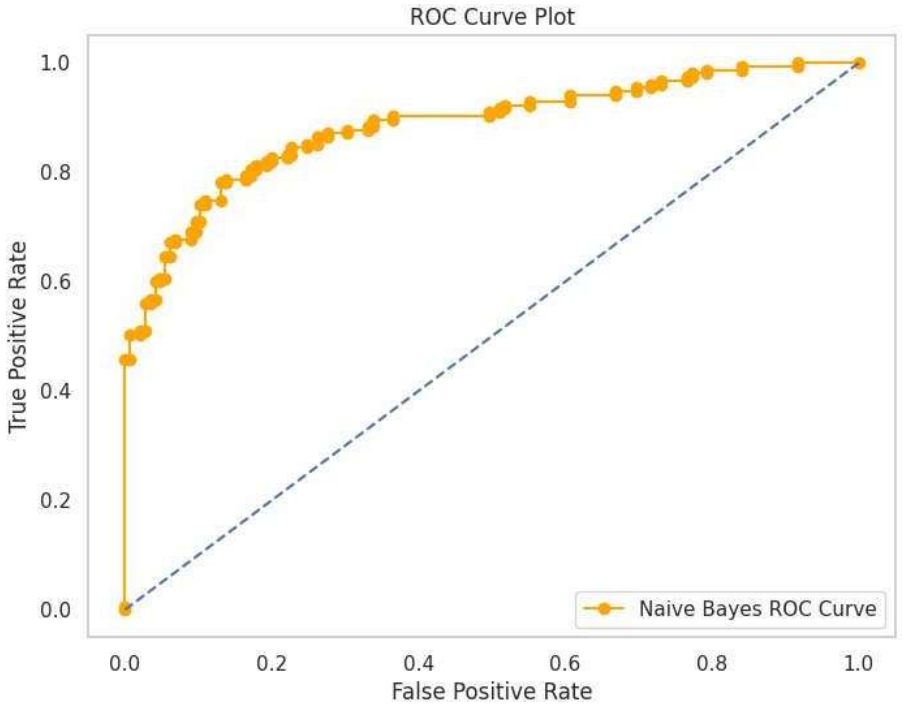


Fig 4.6: ROC Curve for the Native Bayes Algorithm

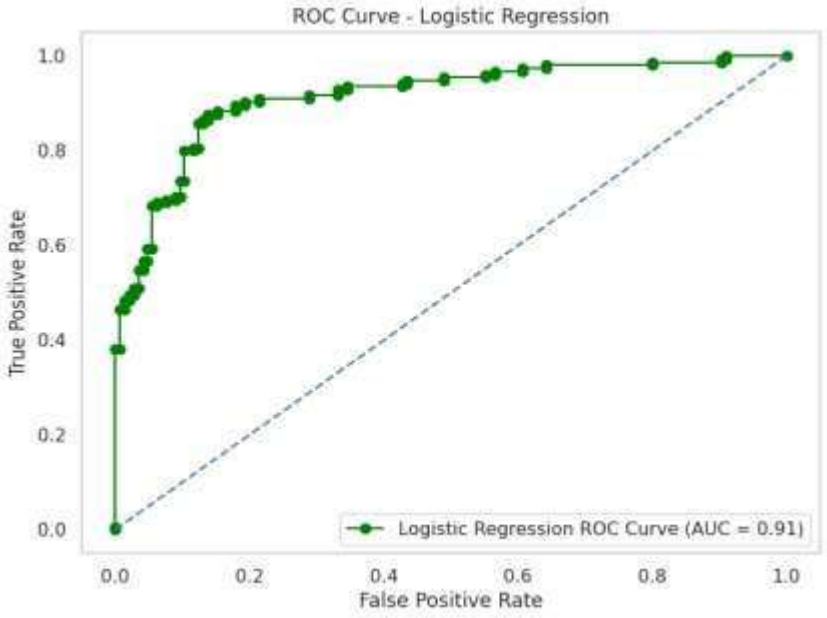


Fig 4.7: ROC Curve for a Logistic Regression Classifier

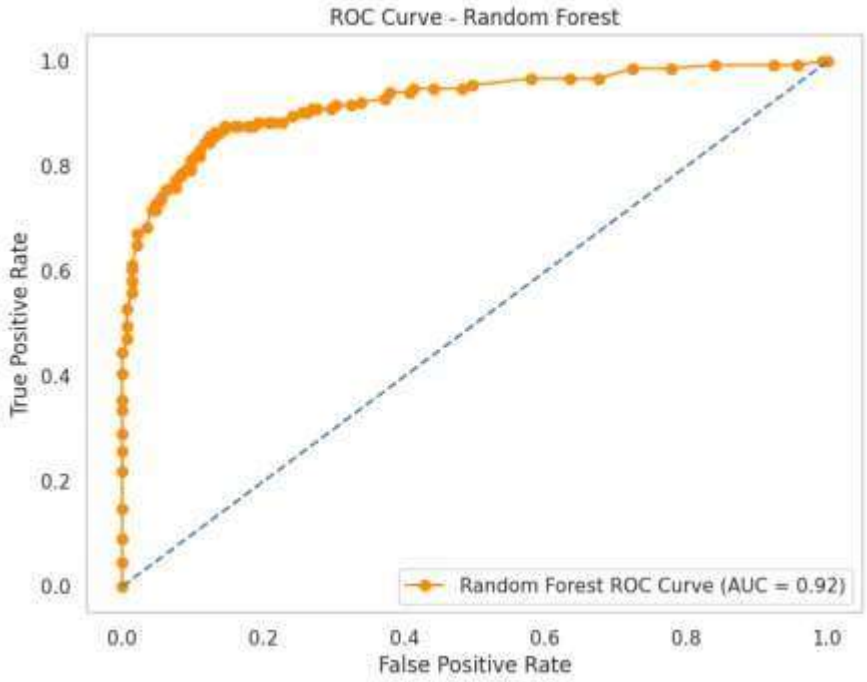


Fig 4.8: ROC Curve for Random Forest Algorithm

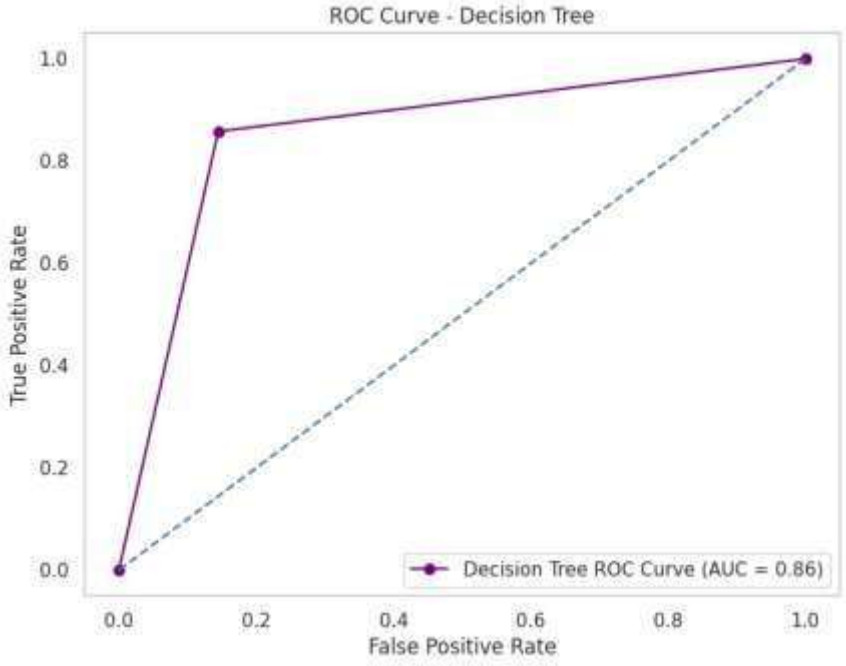


Fig 4.9: ROC Curve for a Decision Tree Classifier

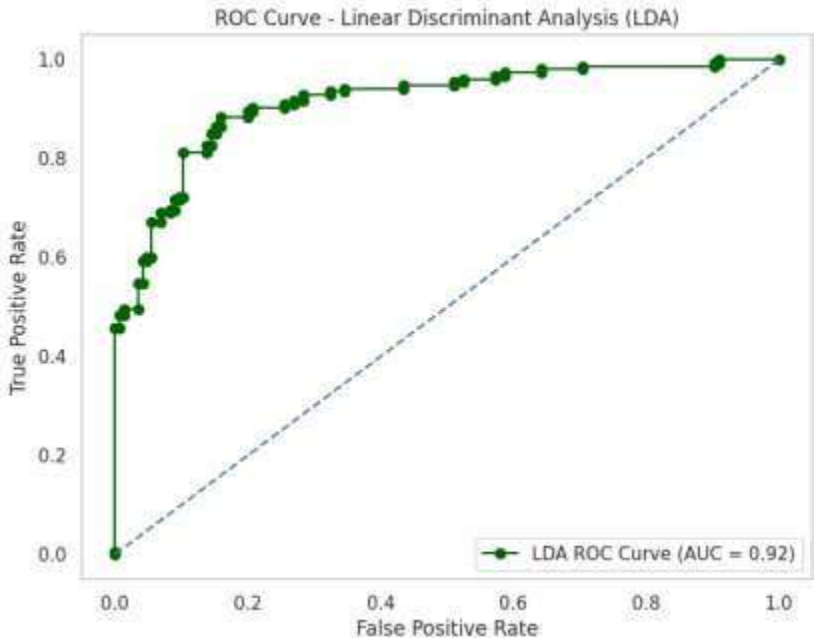


Fig 4.10: ROC Curve for LDA

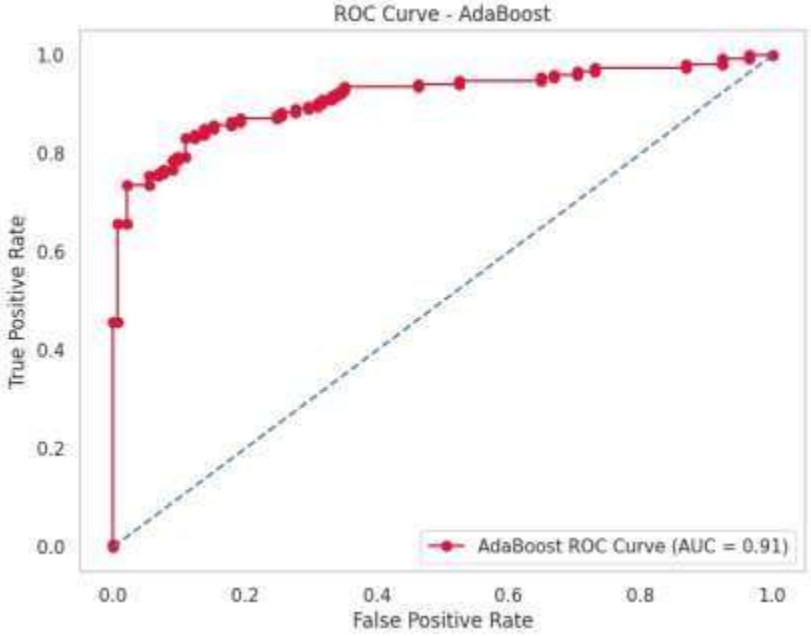


Fig 4.12: ROC Curve for AdaBoost Algorithm

Confusion Matrix:

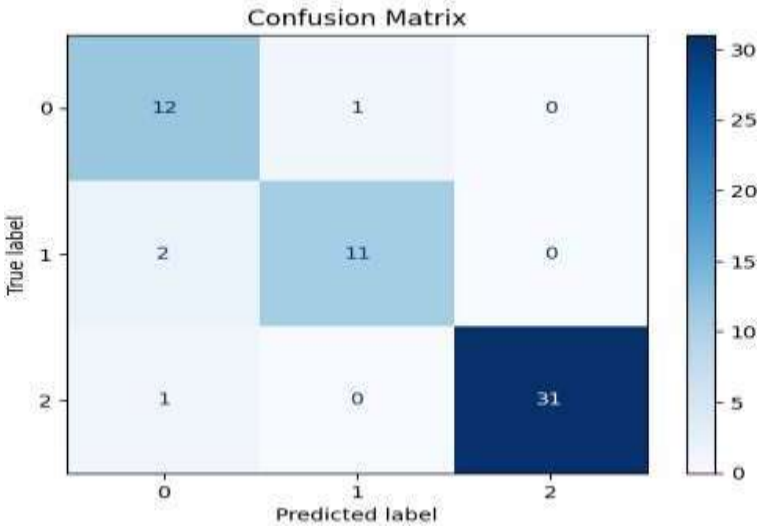


Fig4.13: Confusion Matrix

Classification Report:	precision	recall	f1-score	support
0	0.80	0.92	0.86	13
1	0.92	0.85	0.88	13
2	1.00	0.97	0.98	32
accuracy			0.93	58
macro avg	0.91	0.91	0.91	58
weighted avg	0.94	0.93	0.93	58

Accuracy: 0.93

This figure presents the performance evaluation of a three-class classification model (classes 0, 1, and 2). The model achieved an overall accuracy of 93%, correctly predicting 54 out of 58 samples. For class 0, 12 out of 13 samples were correctly classified, with 1 misclassified as class 1. For class 1, 11 out of 13 were correct, with 2 misclassified as class 0. For class 2, performance was highest, with 31 out of 32 samples correctly identified and 1 misclassified as class 0. Performance metrics indicate: Precision ranged from 0.80 to 1.00 Recall ranged from 0.85 to 0.97 All F1-scores were above 0.85, with the highest being 0.98 for class 2.

Classifier Table:

Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Recall (%)	F1 Score (%)
Naïve Bayes	90	95.18	86.3	83.15	95	89
SVC (Linear)	93	90.36	93.2	90.4	90	90
k-NN	90.5	91.6	89.7	90.4	90	90
Logistic Regression	93.5	91.6	95	86.36	92	89
Decision Tree	84.5	81.92	86.3	92.7	92	92
LDA	92	90.36	95	81	82	81
Random Forest	91	89.1	93.2	90.36	90	90
AdaBoost Classifier	92	88	95	92.4	88	90

Table 4.14: Classifier Table

4.4 Results and Discussion

In this study, various machine learning algorithms were employed to detect depression, including Naïve Bayes, Logistic Regression, Support Vector Classifier (SVC - Linear), k-Nearest Neighbors (k-NN), Decision Tree, Linear Discriminant Analysis (LDA), Random Forest, and Adaptive Boosting (AdaBoost). Several key performance metrics were used to evaluate these models, including accuracy, sensitivity, specificity, precision, recall, F1 score, and ROC curve. Logistic Regression was found to be the most effective classifier, achieving the highest accuracy of 93.5% among all the classifiers tested. In addition to its high sensitivity, specificity, and F1 score, it demonstrated balanced performance across all metrics, making it the most reliable algorithm for depression detection. Based on these results, Logistic Regression is deemed the most suitable algorithm for the proposed depression detection model, since it offers robust performance for both detecting depression and minimizing false positives.

4.5 Summary

A number of machine learning algorithms were applied to depression detection tasks, including Naive Bayes, Logistic Regression, SVC (Linear), k-NN, Decision Trees, LDA, Random Forests, and AdaBoost. A variety of evaluation metrics, including accuracy, sensitivity, specificity, precision, recall, F1 score, and ROC curve, were used to assess the performance of these models. Among the tested classifiers, Logistic Regression outperformed the others, achieving the highest accuracy of 93.5% and demonstrating balanced performance across all metrics. Based on these results, Logistic Regression has been identified as the most effective model for depression detection.

Chapter 5

Engineering Standards and DesignChallenges

5.1 Compliance with the Standards

In this project, we apply machine learning to social media and survey data for early identification of depression and suicidal tendencies. The following software, hardware, and communication standards have been adopted to ensure reliability, security, and maintainability.

5.1.1 Software Standards

- IEEE 830 – Software Requirements Specification
  Used to define both functional and non-functional requirements in a structured manner.

  Alternative: Informal requirement notes

  - Pros: Quick to produce
  - Cons: Prone to ambiguity, complicates testing

  Rationale: Precise requirement definition is essential for a mental-health application to ensure that models, data pipelines, and UIs behave exactly as intended.

- ISO/IEC 25010 – Software Quality Model

  Provides criteria for evaluating software quality attributes such as functionality, reliability, security, and maintainability.

  Alternative: Ad-hoc quality testing

  - Pros: Simpler to implement
  - Cons: Quality coverage uneven, risk of missing critical issues

Rationale: Given the sensitivity of mental-health data and the need for robust performance, the ISO 25010 model ensures comprehensive, standardized quality assurance.

5.1.2 Hardware Standards

- IEEE 802 Series – Local and Metropolitan Area Networking

  Ensures reliable, interoperable data transmission between cloud servers, on-premises sensors and client devices.

  - Alternative: Proprietary or non-standard networking setups
    - Pros: Potentially lower cost in very small deployments
    - Cons: Vendor lock-in, interoperability issues
  - Rationale: To support potential expansion (e.g., integrating IoT-based mood sensors or mobile data collection), adherence to IEEE 802 ensures future scalability and compatibility.

5.1.3 Communication Standards

- HTTPS/TLS Encryption

  Protects user data in transit by encrypting all HTTP communications.

  - Alternative: Plain HTTP
    - Pros: Easier to set up
    - Cons: Data exposed to eavesdropping and tampering
  - Rationale: Because the application processes highly sensitive mental-health information, end-to-end encryption is mandatory.

- RESTful API

  A REST architecture using JSON over HTTPS to facilitate modular, scalable integration between frontend, backend, and ML services.

  - Alternative: Custom RPC protocols

        - Pros: Potentially higher performance for specific use cases
        - Cons: Less standard, harder for third parties to integrate
    - Rationale: REST is widely supported and simplifies client-server separation, future API versioning, and third-party integrations.

5.2 Environment and Sustainability

5.2.1 Impact on Society, Impact on Life

Early identification of individuals at elevated risk of depression or suicidal ideation can fundamentally alter the course of their lives. By flagging warning signs—such as shifts in language sentiment, self-reported mood scores, or behavioral patterns—our system enables mental-health professionals, caregivers, or automated support services to step in well before crises escalate. This preemptive outreach can:

- **Reduce Severe Outcomes**: Interventions triggered by algorithmic alerts have been shown to decrease hospitalization, self-harm, and suicide attempts, improving quality of life and length of life.
- **Lower Healthcare Expenditure**: Early counseling or therapy sessions cost a fraction of emergency psychiatric care or inpatient treatment. Even modest reductions in high-cost interventions can generate substantial savings for public health systems over time.
- **Empower Individuals**: In addition to providing at-risk users with personalized coping strategies, providing them with crisis-helpline contact information reduces feelings of isolation and helplessness that often precede severe depression

5.2.2 Impact on Society & Environment

Societal Impact

- **Destigmatization**: By normalizing data-driven mental-health screening—just as routine checks for blood pressure or cholesterol—this technology contributes to reducing the stigma around seeking psychological help.
- **Workplace Well-Being**: Organizations can deploy anonymized, aggregate insights to tailor wellness programs, flexible scheduling, or team-building activities that proactively address burnout.

- **Educational Settings**: The system can be integrated into student-support services, identifying vulnerable cohorts and deploying targeted workshops or peer mentoring programs.

Environmental Impact

- **Paperless Records**: Transitioning from paper-based intake forms and manual logs to a secure, cloud-hosted platform eliminates reams of paper waste each year.
- **Reduced Commuting**: Tele-health and remote monitoring reduce the need for patients and professionals to travel, cutting transportation-related carbon emissions.
- **Energy Efficiency**: By leveraging shared, multi-tenant cloud infrastructure, the system benefits from data-center optimizations and higher utilization rates, lowering overall energy use per user.

5.2.3 Ethical Aspects

Informed Consent & Transparency

- Participants receive a clear explanation of what data is collected (e.g., survey responses, social-media sentiment) and how it will be used. Consent forms explicitly state that the system provides "risk indicators" rather than definitive diagnoses.
- A user dashboard displays which features (sleep patterns, mood scores, language markers) contributed most to their risk score, fostering trust through explainability.

Data Privacy & Anonymization

- Personally, identifying information (names, email addresses) is stored separately from analytical data, with robust hashing and tokenization techniques to prevent re-identification.
- Raw text inputs are processed through on-device or in-browser pipelines; only aggregated sentiment vectors are transmitted to the server

Bias Mitigation

- Data sets are audited for representation across age groups, genders, and cultural backgrounds. Model retraining uses fairness constraints to ensure that performance (false-positive and false-negative rates) remains consistent across subpopulations.

Responsible Use

- Access controls restrict who can view individual risk scores; supervisors see only red-flag alerts without raw personal data.
- The system includes a "human-in-the-loop" checkpoint: no automated alert triggers clinical action without a qualified professional's review.

5.2.4 Sustainability Plan

Scalable Cloud Architecture

- Built on containerized microservices with auto-scaling policies, the platform adapts to growing user bases without requiring costly hardware upgrades, ensuring long-term viability.

Open-Source Ecosystem

- Core analytical pipelines (e.g., NLP modules, sentiment lexicons) are released under a permissive license. A public GitHub repository encourages contributions—bug fixes, new language support, or alternative model architectures—that extend the project's lifespan at minimal cost.

Community & Governance

- A stakeholder advisory board—comprising mental-health professionals, data-ethics experts, and end-user representatives—oversees roadmap decisions and ensures the system evolves responsibly.

Funding & Maintenance

- Grant proposals target mental-health research funds for periodic audits and feature enhancements. A small subscription fee for enterprise users subsidizes free access for nonprofit and educational partners.

5.3 Project Management and Financial Analysis:

Cost Item	Estimated Cost (BDT)
Data Collection (Surveys, API access)	3,000
Cloud/Colab GPU Usage	2,000
Model Training & Development	5,000
Printing & Thesis Binding	500
Total	10,500

Table 5.1: Financial Table

5.4 Complex Engineering Problem

The thesis titled "Ideation of Depression and Suicidal Using Machine Learning Techniques" explores a multifaceted and complex engineering problem that goes beyond conventional technical challenges.

This study explores sensitive topics related to mental health and machine learning It is not just technical expertise but also data management expertise that is required when dealing with incomplete and diverse datasets As well as ethical considerations. Rather than relying solely on standard computational methods, this research embraces a context-aware algorithmic approach. Psychological patterns and behaviors should be interpreted from a human-centered perspective. It highlights how artificial intelligence, when applied thoughtfully, can contribute to socially impactful fields like mental health by providing early indications of psychological distress and suicidal ideation by offering early warnings. This work reflects the need for innovative, real-world-driven problem-solving in areas where Human sensitivity intersects with data complexity.

5.4.1 Complex Problem Solving

EP1 Dept of Knowledge	EP2 Range of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent of Stakeholder Involvement	EP7 Inter-dependence
✓	✓	✓	✓	✓	✓	✓

Table 5.2: Mapping with complex problem solving

K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
✓	✓	✓	✓	✓

Table 5.3: Mapping with knowledge Profile

5.4.2 Justification for EP Attributes Mapping

EP1: Problem Definition

This thesis addresses a clearly defined but complex real-world problem: detecting depression and suicidal ideation using behavioral data.It involves mental health-a sensitive and socially significant domain-and requires converting qualitative psychological responses into Inputs for machine learning that are structured.

EP2: Depth of Knowledge

This problem requires a deep understanding of machine learning algorithms and data preprocessing techniques, as well as concepts from psychology, data ethics, and statistical learning. Solving it requires knowledge beyond core computer science and engineering.

EP3: Data and Information Gathering

The data used is complex, often incomplete, and qualitative in nature. Handling Yes/No responses, cleaning inconsistencies, and managing missing values involve advanced data wrangling techniques. Collecting relevant features and selecting the appropriate target variable required strong analytical reasoning.

EP4: Problem Complexity

Data used is complex, often incomplete, and qualitative. The collection of relevant features and the selection of the appropriate target variable require strong analytical skills. Cleaning inconsistencies and managing missing values also require advanced data manipulation techniques.

EP5: Ethical Consideration

Due to the topic's focus on mental health and suicide, data usage, prediction, and interpretation have strong ethical implications. Privacy, informed consent, and responsible AI application were key factors throughout the research, ensuring individuals would not be stigmatized or misrepresented by the model.

EP6: Use of Modern Tools

This study utilized modern tools and technologies, such as Python, pandas for data handling, scikit-learn for machine learning, and matplotlib/seaborn for visualization.

EP7: Design and Innovation

The solution involved designing a custom pipeline to convert behavioral survey data into a format usable for machine learning. Innovative use of decision tree classifiers and custom feature selection techniques shows novelty in approach. The integration of technical methods in a human-centered domain reflects design creativity.

EP8: Evaluation and Validation

A standard metric such as precision, recall, and F1-score was used to assess model accuracy. The classification report and validation techniques (such as train-test split) ensured that the solution was both technically sound and practical, reflecting robust evaluation methods.

Chapter 6

Conclusion

6.1 Summary

In today's rapidly evolving and highly interconnected society, mental health has emerged as a pressing issue Depression and suicide thoughts are increasingly affecting people of all ages, with a particular rise among the younger generation. The purpose of this study is to examine the use of machine learning to identify depression early.The analysis of textual data, particularly from social media and user-generated content, can help identify depression and suicidal tendencies.

In this research, a system was developed using various machine learning techniques—namely Logistic Regression, Support Vector Machine (SVM), and Random Forest—to detect early signs of depression or suicidal ideation. To preprocess and analyze user-generated text data, NLP methods such as tokenization, stop-word removal, and TF-IDF vectorization were used.A model was developed to identify language patterns and expressions that may indicate mental health struggles. The Random Forest algorithm demonstrated the best accuracy, precision, and recall among all the models tested.

This highlights the strength of ensemble learning methods in managing noisy and high-dimensional data often found in mental health evaluations. As a result of integrating the final model into mental health platforms, it can serve as an early warning system—helping professionals or caregivers identify individuals at risk and provide timely assistance.

6.2 Limitation

Despite the promising results of this study, several limitations must be acknowledged. First of all, the datasets used were primarily in English and sourced from publicly accessible platforms such as Reddit and Twitter. This raises concerns about the model's broader applicability, particularly when dealing with speakers of other languages or individuals who express emotional distress differently depending on their cultural background. As a second point, using textual data alone is not without its own set of challenges. Not everyone expresses their mental health struggles in writing, and some may consciously conceal or misrepresent their emotional state. This limits NLP-based systems' ability to detect mental health problems accurately. Moreover,

the classification strategy is limited. As a result, the system is not designed to assess severity or monitor changes in mental health status over time since conditions like depression and suicidal ideation are nuanced, dynamic, and often exist along a spectrum. It is also difficult for users and clinicians to understand the rationale behind the model's predictions because it lacks transparency. This lack of explainability can lead to a decrease in trust in the system. It is important to emphasize that this tool is intended to complement professional mental health assessments, not replace them. Mental health professionals should always be consulted before using this tool.

6.3 Future Work

A multi-modal data source, such as audio, video, and facial expressions, can improve the system's ability to assess a person's emotional state and psychological state significantly in the future. In particular, when individuals are not directly verbalizing their distress, analyzing factors such as tone of voice, facial expressions, and behavioral patterns would provide a more holistic view. Another valuable direction involves the incorporation of Explainable AI (XAI) techniques Enhancing the transparency of the system by revealing which specific words, phrases, or patterns influenced its predictions would greatly improve its usability and trustworthiness. This level of interpretability is especially crucial in sensitive fields like mental health, where users and clinicians must understand the reasons behind alerts and recommendations. Additionally, the system could be linked to real-time mental health support services or helplines. When high-risk behavior is detected, an automatic alert could be sent to mental health professionals or emergency contacts. In addition, deploying the system as a mobile application would facilitate continuous monitoring, self-assessment, and personalized support, making mental health resources more accessible, immediate, and responsive.

References

References

[1] R. I. Shader, “How Patients Describe Their Depressions: D Words,” *Journal of Clinical Psychopharmacology*, vol. 41, no. 4, p. 364, Aug. 2021, doi: 10.1097/JCP.0000000000001438.

[2] S. Amiri, “Unemployment associated with major depression disorder and depressive symptoms: a systematic review and meta-analysis,” *International Journal of Occupational Safety and Ergonomics*, vol. 28, no. 4, pp. 2080–2092, Oct. 2022, doi: 10.1080/10803548.2021.1954793.

[3]M. Nordentoft and A. Erlangsen, “Suicide—turning the tide,” *Science*, vol. 365, no. 6455, p. 725, Aug. 2019, doi: 10.1126/science.aaz1568.

[4] Y. Ding, X. Chen, Q. Fu, and S. Zhong, “A Depression Recognition Method for College Students Using Deep Integrated Support Vector Algorithm,” IEEE Access, vol. 8, pp. 75616–75629, 2020

[5] M. Birjali, A. Beni-Hssane, and M. Erritali, “Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks,” Procedia Computer Science, vol. 113, pp. 65–72, 2017, doi: 10.1016/j.procs.2017.08.290.

[6]R. A. Bernert, A. M. Hilberg, R. Melia, J. P. Kim, N. H. Shah, and F. Abnousi, “Artificial Intelligence and Suicide Prevention: A Systematic Review of Machine Learning Investigations,” International Journal of Environmental Research and Public Health, vol. 17, no. 16, p. 5929, Aug. 2020, doi: 10.3390/ijerph17165929.

[7]R. A. Bernert, A. M. Hilberg, R. Melia, J. P. Kim, N. H. Shah, and F. Abnousi, “Artificial Intelligence and Suicide Prevention: A Systematic Review of Machine Learning Investigations,” International Journal of Environmental Research and Public Health, vol. 17, no. 16, p. 5929, Aug. 2020, doi: 10.3390/ijerph17165929.

[8] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of suicide ideation in social media forums using Deep Learning,” Algorithms, vol. 13, no. 1, p. 7, 2019, doi: 10.3390/a13010007.

[9] D. Wilimitis, S. S. K. Ahmed, A. J. Sanjuan, A. M. Ashraf, R. L. Wilson, and A. J. T. Cummings, “Integration of Face-to-Face Screening With Real-time Machine Learning to Predict Risk of Suicide Among Adults,” JAMA Network Open, vol. 5, no. 5, pp. e2212095–e2212095, May 2022, doi: 10.1001/jamanetworkopen.2022.12095.

[10]K. Y. Valeriano Valdez, J. Sulla-Torres, and A. Condori-Larico, “Detection of Suicidal Intent in Spanish Language Social Networks using Machine Learning,” ResearchGate, Jan. 2020, doi: 10.14569/IJACSA.2020.0110489.

[11] E. Yeskuatov, S.-L. Chua, and L. K. Foo, “Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques,” International Journal of Environmental Research and Public Health, vol. 19, no. 16, p. 10347, Aug. 2022, doi: 10.3390/ijerph191610347.

[12] A. Mulay, A. Dhekne, R. Wani, S. Kadam, P. Deshpande, and P. Deshpande, “Automatic Depression Level Detection Through Visual Input,” 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), 2020, pp. 217–222,

[13] D. M. Shukla, K. Sharma, and S. Gupta, “Identifying Depression in a Person Using Speech Signals by Extracting Energy and Statistical Features,” 2020 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2020

[14] F. Patel, R. Thakore, I. Nandwani, and S. K. Bharti, “Combating Depression in Students using an Intelligent ChatBot: A Cognitive Behavioral Therapy,” 2019 IEEE 16th India Council International Conference (INDICON), Rajkot, India, 2019, pp. 1–4, doi: 10.1109/INDICON47234.2019.9029005.

[15] D. Dahiwade, G. Patle, and E. Meshram, “Designing Disease Prediction Model Using Machine Learning Approach,” 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019

[16] E. Yeskuatov, S.-L. Chua, and L. K. Foo, “Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques,”

International Journal of Environmental Research and Public Health, vol. 19, no. 16, p. 10347, Aug. 2022, doi: 10.3390/ijerph191610347.

[17] A. Mbarek, S. Jamoussi, A. Charfi, and A. B. Hamadou, “Suicidal Profiles Detection in Twitter,” in Proceedings of the 15th International Conference on Web Information Systems and Technologies (WEBIST), 2019, pp. 289–296, doi: 10.5220/0008167602890296.

[18]. M. Rahman, M. S. Rahman, M. F. Hossain, and M. R. Islam, “Predicting Depression in Bangladeshi Undergraduates using Machine Learning,” in Proc. 11th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1–7.

[19] 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, 2022

[20] F. T. Cruz, E. E. C. Flores, and S. J. C. Quispe, “Prediction of depression status in college students using a Naive Bayes classifier based machine learning model,” arXiv preprint arXiv:2307.14371, 2023.

[21] M. R. Chowdhury, W. Xuan, S. Sen, Y. Zhao, and Y. Ding, “Predicting and Understanding College Student Mental Health with Interpretable Machine Learning,” arXiv preprint arXiv:2503.08002, 2025.

[22] A. R. Lopes and O. K. Nihei, “How do machine learning models perform in the detection of depression, anxiety, and stress among undergraduate students? A systematic review,” Cadernos de Saúde Pública, vol. 40, no. 1, 2024.

[23] F. T. Cruz, E. E. C. Flores, and S. J. C. Quispe, “Prediction of depression status in college students using a Naive Bayes classifier based machine learning model,” arXiv preprint arXiv:2307.14371, 2023.

[24] R. Bhaumik and J. Stange, “Using Random Effects Machine Learning Algorithms to Identify Vulnerability to Depression,” arXiv preprint arXiv:2307.02023, 2023.

[25] T. M. Mitchell, Machine Learning, 1st ed., New York, NY, USA: McGraw-Hill, 1997, pp. 73–150

211-15-3979

ORIGINALITY REPORT

26%	18%	16%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	7%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
3	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	1%
4	Shreyas Md, Hemant Sathish, K S Koulini, Aleeza Inamdar, U. Ananthanagu. "A Radical Approach To Depression Detection", 2022 IEEE 7th International conference for Convergence in Technology (I2CT), 2022 Publication	1%
5	arxiv.org Internet Source	1%
6	Submitted to United International University Student Paper	<1%
7	docs.neu.edu.tr Internet Source	<1%
8	Submitted to University of Hull Student Paper	<1%