EARLY DEPRESSION DETECTION USING MACHINE LEARNING: A COMPARATIVE STUDY OF ENSEMBLE MODELS AND FEATURE SELECTION WITH RFE

BY

Farjana Abedin Boby
241-25-046

This Report Presented in Partial Fulfilment of the Requirements for
The Degree of Master of Science in Computer Science and Engineering

Supervised By

Abdus Sattar

Associate Professor & Director MSc Program
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Arif Mahmud

Associate Professor and Associate Head

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

May 2025

APPROVAL

This Thesis titled **"Early Depression Detection Using Machine Learning A Comparative Study of Ensemble Models and Feature Selection with RFE"**, submitted by **Farjana Abedin Boby**, ID No: **241-25-046** 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 **MSc. in Computer Science and Engineering** and approved as to its style and contents. The presentation has been held on **24-05-2025.**

BOARD OF EXAMINERS

Chairman

Dr. Arif Mahmud
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md Alamgir Kabir
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Nazibur Rahman
Technical Lead, Database Administrator
Wipro, Telenor - Grameen Phone Account
Dhaka, Bangladesh

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Abdus Sattar**, Assistant Professor, Department of CSE, Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Abdus Sattar

Associate Professor & Director MSc Program
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Arif Mahmud

Associate Professor and Associate Head
Department of CSE
Daffodil International University

Submitted by:

Farjana

Farjana Abedin Boby

ID: 241-25-046
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to Almighty Allah for His divine blessing which makes it possible to complete the final year project/internship successfully.

I am grateful and wish my profound indebtedness to Abdus Sattar, Assistant Professor, Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest of my supervisor in the field of Machine Learning to 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 them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude Dr. Arif Mahmud, Associate Professor and Associate Head, Department of CSE, for his kind help to finish our project and to other faculty members and the staff of CSE department of Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Depression is a serious mental disease which has affected millions of people all over the world, but it is very difficult to detect in early-stage. Traditional diagnosis is Mesh Keywords based on self-administered questionnaires and clinical diagnostics, and both are subjective and time-consuming. In this work, we use machine learning to predict depression based on a dataset from Kaggle containing multiple behavioral, psychological and demographic variables. The research concentrates on feature selection technique, RFE and evaluates its performance by using various machines' learning models like XGBoost, Gradient Boosting, AdaBoost and Logistic Regression. Much preprocessing was performed on the dataset, such as managing missing values and scaling of features. A set top 10 key features were chosen through RFE for lessening the dataset complexity and for maximizing the predicting the accuracy. Although ensemble models such as XGBoost and Gradient boosting model attained the highest accuracy for all the experiments (95%), testifying that RFE was able to reduce the dimensionality of the dataset without the loss of discriminating power. The findings suggest AI-enabled predictions of mental health may be a scalable and unbiased approach to screening for depression. The above avenues can form the future direction using deep learning architectures, multimodal data sources and privacy preservative strategies for real world application. This paradigm serves in the expansion of the burgeoning area of artificial intelligence in mental health and provides evidence on the effectiveness of machine learning in predicting early identification of depression and algorithm driven interventions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

It is a major mental comorbidity that affects million people worldwide with profound negative impact on their daily functioning. Severe hopelessness and worthlessness, along with an inability to feel any motivation or interest in things that were previously enjoyed are the trademarks of this persistent economic disorder.” It is something more than mere sadness or passing occasional discontent. Depression is one of the greatest causes of disability worldwide with more than 280 million individuals suffering from it in the world according to the World Health Organization (WHO). It’s a multifactorial problem, with biological, psychological, and environmental etiologies. Thoughts of suicide, difficulty concentrating, changes in eating and sleeping patterns, a lack of energy and constant sadness all are symptoms of depression. Reactions of depression can also be characterized as Major Depressive illness (MDD), Seasonal Affective Disorder (SAD), bipolar illness, and Persistent Depressive Disorder (Dysthymia). The illness can range in severity; some have mild symptoms that amount to discomfort, while others have severe attacks that need medical intervention. Depression is also common; not a lot of people know about it, and stigma, misinformation or a lack of access to mental health services prevent many from seeking professional help. Depression is a core feature of mental health and constitutes one of the most frequent psychiatric diagnoses. It is often comorbid, such as with anxiety disorders, bipolar disorder, and schizophrenia. Depression and other mental illnesses can be hard to diagnose and treat, because symptoms can sometimes overlap. Neurobiological studies are also pointing to chemical imbalances in the brain—specifically neurotransmitters such as serotonin, dopamine and norepinephrine that help regulate mood and manage emotions—as a likely cause of sadness. There are psychological factors such as isolation, trauma, loss, and persistent stress that contribute to depression in addition to physical causes. Personal problems, financial problems, academic stress, and occupational stress are other important determinants of depression. The COVID-19, a terrible global pandemic that created an unprecedented rise in anxiety and hopelessness especially in this era of global lockdowns, economic instability and uncertainty, further affirmed the importance of mental health. Even though society has acknowledged the role of an illness, mental health therapy in most of the world is still in the embryonic stage. Valuable mental health resources are scarce and stigma surrounding psychiatric illness inhibits many from seeking and accessing the care they need in low- and middle-income countries (LMICs). Mental health services are also frequently hard to find and expensive even in wealthy countries, impairing the ability to treat depression on a large scale. "Depression is prevalent among the young and academic population, workforce starters in particular. Youth (15–29-year-old) are especially at risk of depression so much due to social pressure, peer pressure, career uncertainty, and academic stress. Research indicates that suicide is the second leading cause of death for young people, and is frequently associated with untreated depression. In the school setting, depression can have a deleterious effect on motivation, cognition, and academic functioning. For students with mental health conditions, difficulties in focusing, lack of interest in school work and social withdrawal are common barriers that negatively impact on academic achievement. There is even a level where Welfare and dropout rates rise as students are not able to weather those chronic stress and worry storms. Social media and digital platforms can also be blamed for the rise in youth depression cases. Even factoring in unreasonable social expectations, cyberbullying and exposure to romanticized versions of life, all-day screen time cannot be healthy for self-image and probably causes real mental anguish. Young people more depressed: Modern lifestyle AND ever younger the increase in the depression of young people has also been linked to the modern lifestyle: bad eating habits, little sleep, little exercise. It is important to combat depression in the young, intervention early in life has profound effects on morbidity and can increase quality of life, productivity and career success. Schools, organizations and companies should build mental illness awareness programs and support networks on campuses and in the workplace to help people notice symptoms and seek professional help. 2 ©Daffodil International University As depression has an enormous public health impact, there are many reasons to investigate it. By diagnsoing and troptic depression in it’s early stages we can help prevent the long-term effects of untreated depression such as drug addictiob, alcoholism, sucide, and self-hame. Still, most traditional clinical tools for diagnosing depression are based on self-reported questionnaires and psychiatry interviews and can be subjective and subject to underreporting because of social stigma. Recent developments in the field of artificial intelligence (AI) and machine learning (ML) makes it possible to build predictive and diagnostic models for depression. Petabytes When analysed by machine learning algorithms, physiological signals, data from wearables, voice and text analysis, along with social media activity and behavioural patterns can lead to more accurate, and objective diagnosis of depression. Leveraging large datasets, such algorithms tap into underlying patterns that might evade traditional psychiatric assessments. In addition, machine learning-mental health prediction systems might provide scalable and cheap options for mental health screening. They may be incorporated into chatbots, mobile phone apps, and wearable devices to allow greater public access to mental health monitoring. This research is essential to lowering the global burden of mental health difficulties, such as stigmatization, burden of care and cost associated with mental health, early intervention, and prediction accuracy. Beyond the impact on people, depression has a significant social and economic impact. Depression can lead to work absenteeism, decreased productivity, higher health-related costs, and even death. Mental health disorders not only reduce worker productivity, they also cost the global economy billions of dollars each year, studies show. And if untreated depression among workers results in difficulty making decisions, building relationships in the workplace and completing tasks, that could be bad for business.

1.2 Motivation

The availability of depression, particularly to young people, is increasingly recognised as a major public health issue. This cause for concern underscores the urgent need for better, more efficient, and more accessible mental health diagnostic tools. Although effective, traditional clinical methods have several limitations. These methods are generally over-dependent on subconscious self-reporting which can be more or less unreliable and do not necessarily reflect the true mental state of a particular individual– they are also invariably expensive and time-consuming. Thus, the risk of late diagnosis is increased being that it could accelerate the progress of this kind of mental illness, and lead to awful consequences, such as suicide. Amid nucleation in a data-driven world of pervasive digital connectivity, machine learning provides an unprecedented opportunity to advance mental health diagnostics. The capability of AI algorithm to analyze huge and complex data-sets, which can easily overlook subtle trends and early warning signs not readily available to standard clinical examinations. One of the key drivers for this work is the potential for AI to offer rapid, cost-effective and scalable diagnostic assistance.

Leveraging the machine learning techniques for more accurate and efficient prediction on depression is one of our primary research interests. Among them, Recursive Feature Elimination (RFE) has been employed to refine the dataset by selecting the significant features. This procedure will reduce the complexity of computation and enhance the prediction performance of the model by omitting redundant or useless contents. This work is motivated not only by technological advances, but also by a significant social and personal imperative: to reduce the stigma surrounding mental health and to call for data-driven, easily accessible solutions. Depression affects the whole person; from one's physical health, to one's mood, and to one's relationship with others, work and academic outcomes. We aim to empower people and patients as well as healthcare systems to better recognize and find the right level of care in a more efficient proactive way leveraging user-friendly widely accessible AI-powered screening tools.

1.3 Rationale of the Study

Depression is a public health problem with increasing financial, social, and psychological costs. Although widely used, the early identification is a significant barrier as it depends on conventional diagnostic means such as clinical interviews and self-report questionnaires that can be time-consuming, subjective or unavailable periodically. These restrictions delay care and can worsen symptoms and increase the risk of major outcomes such as suicide and chronic mental illness. In the past few years, machine learning (ML) which provides a more objective, data-driven approach is showing promising hope for revolutionizing the diagnosis of mental health. However, most related works concentrate on a small number of features, or they do not well exploit feature selection to enhance model performance. The need for valid, scalable, and efficient instruments for mental health screening that can be embedded in digital platforms justifies this work. By reducing the dimension of the dataset and increasing the accuracy of the model, the paper hopes to contribute to the development of intelligent systems for the detection and early prevention of depression that could help alleviate the global burden of the disease.

1.4 Research Questions

The following three research questions constitute the foundation of this investigation:

1. Some data preprocessing techniques such as dealing with missing values, and converting string data to numerical data help make a depression dataset from kaggle better?
2. How much does the Recursive Feature Elimination (RFE) reduce model complexity and select proper features to predict depression?
3. What are the predictive implications for depression when using and not using RFE for several ML models (Logistic Regression, AdaBoost, Gradient-Boosting and XGBoost)?

1.5 Expected Output

The main purpose of this study is to analyze and predict depression symptoms using machine learning algorithms based on a Kaggle data set. Depression is a severe mental illness that affects millions of people worldwide, and timely intervention and treatment depend on early recognition. Psychiatric evaluation, self-report questionnaires, and subjective and laborious clinical assessments are traditional methods of diagnosing depression. However, with the development of machine learning and data-driven technique, automated models predicting depression have emerged and can evaluate a large variety of features and identify trends that one cannot see with conventional methods.

The following leaf is employed to achieve the aims of this work in this systematized machine learning pipeline: 1.

- Data collection and preprocessing: importing the dataset from Kaggle and using EDA to handle any missing values, encode categorical variables and understand the distribution of the data.
- Recursive Feature Elimination (RFE) for Feature Selection: Ask for what are the variables responsible for depression.
- Model Application and Performance Evaluation: We compare model predictive performance with and without RFE in training and testing different ML algorithms. The most effective strategy is determined by the model accuracy, precision, recall, and F1-score, as well as the confusion matrices.

We hope to improve the predictive model of depression by the discovery of significant risk factors for mental health disorder. The findings could help enhance systems that diagnose mental health problems, so professionals can use data to decide how to intervene in a patient's life.

1.6 Project and Finance Management

This study is not funded or managed. It was not sponsored by external sources, institutional grants or third-party funds and was performed as a requirement of the course. The entire research pipeline, including data collection and pre-processing, model design, testing and documentation, was carried out by in-house resources. Project management applied the systematic and time-limited process to ensure good progress. From the start of the project, I made a project plan, dividing the job into high-level phases: the literature review, the exploration of data sets, the feature engineering, the model implementation, the result analysis, and the writing of the thesis. We stayed organized and on course with tools like Trello for tracking our tasks, and Git for versioning. It was free to the public from Kaggle so no data collecting costs were incurred. Experimentation and visualisation was performed with Python based libraries such as Pandas, Scikitlearn, XGBoost and Matplotlib in Jupyter Notebook, with no extra charge. Paid software or any cloud computing were not required, as all the analyses were conducted on a personal computer.

1.7 Report Layout

This thesis's architecture offers a general-purpose framework for depression prediction via machine learning. Through a systematic model-centric process from data collection toward model comparison, it ensures a comprehensive grasp of how machine learning models can predict mental health. Significance and innovate of this paper as below: Kaggle has a dataset of mental health data as keywork, start from Kaggle; Step 1: Data collection; Step 2: Research background and literature review. The dataset contains various factors about a person which are considered to influence the likelihood of being depressive. It is the basis for models for machine learning that classify individuals as depressed or not. Once the data is collected, The research moves on to the next stage of data analysis and visualization to gain insights from the dataset. Diverse statistical and graphic methods are applied to help understanding the distributions of features, the patterns and relationships among variables. Visualizations of depression and its contributors are histograms, box plots, pie charts, and correlation heatmaps. This is also the case for the imbalance class problem, due to the extremely high proportion of not depressed (0) compared to the depressed (1). Data preprocessing is the next step which is done to clean the dataset and prepare it for input into machine learning. This step relates to encoding of categorical variables, imputation of missing values, and scaling of the numeric features, because the class imbalance influences the accuracy of predicting. Preprocessing is a risk mitigation factor, as long as the models can be trained on a balanced and well managed dataset, to prevent biased predictions. After data preparation, feature engineering is applied to detect the important predictors of depression. Using Recursive Feature Elimination (RFE) to evaluate which are most valuable predictors of depression, the model can zero in on the key traits and discard those that are irrelevant. Thus, the identified features may be employed to train comparative models to further investigate performance enhancement through feature selection. The final part of the study is model comparison, using Logistic Regression, Random Forest, XGBoost, Gradient Boosting and AdaBoost to train and test machine learning models. These models are then studied with and without using RFE, and evaluated with key indicators like accuracy, precision, recall, F1score and confusion matrices. Train validation loss plots are also provided to see how well the learning generalizes or overfits. In terms of performance and interpretability, this comparison aids in the selection of an optimal model to predict depression. Overall, this thesis offers well-organized summary of data collection, pre-processing, feature selection, and machine learning models applied to analyze depression. The results assist in the design of AI-based tools for mental health assessment and provide insights into how machine learning could enhance these tools.

CHAPTER 2
BACKGROUND

2.1 Introduction

Mental health is an important component of the well-being of any individual as it affects cognitive capabilities, emotional stability and general quality of life. Depression is a common and disabling mental illness that affects millions of people worldwide. Depression is believed to affect approximately 280 million people worldwide by the World Health Organization (WHO) and is a major contributor to the worldwide burden of diseases and is the leading cause of disability. Depression is not simply a short-lived sadness but is long term sadness stuck where ever it falls for you, is accompanied by boredom, tiredness, trouble concentrating and at its worst suicidal thoughts. The potential physical, emotional and social consequences of depression left untreated can be serious, from losing a job to social isolation, or a greater risk of developing other health problems such as heart disease. The global issue of depression has worsened due to various socio-economic, cultural and technological changes. Increasing financial insecurity, social fragmentation, heightened workplace pressures, and accelerating patterns of urban living have contributed to this epidemic of poor mental health across the board. With lockdowns, economic instability and an uncertain future, studies showed an alarming increase in depression and anxiety and just made the mental health situation worse. What's more, people working from home and communicating online has shifted social dynamics, reduced human interaction and worsened a feeling of isolation. The young and old are particularly vulnerable, with young adults subject to greater academic and social stress and older people often affected by loneliness and failing health. Notwithstanding its high prevalence, especially in low and middle-income countries (LMICs) where access to mental healthcare is scarce, depression is underdiagnosed and undertreated. Because of widespread misconceptions and stigma about mental health, many people who might benefit from care don’t pursue it, and their symptoms can become more severe as the years pass. Furthermore, there is a cultural difference in attitudes to depression and its treatment. Depression is a good example.T Here are cultures that don’t allow you to see a doctor because they believe that depression is a fleeting emotion, not a serious illness. Ignorance about mental health issues adds fuel to the fire, as does poor provisioning of mental healthcare. Technology has been increasingly recognised as potential support in mental health screening and intervention, and artificial intelligence (AI) in particular has progressively become a promising solution to address these challenges. By observing behavioral habits, analyzing the text, speech recognition and social media usage, Machine Learning models have started predicting depression now. Such technologies provide a scalable and economical way to identify individuals at risk for depression, and to monitor and intervene proactively. #x201C; Integration of wearable sensors, smartphone apps, and AI enabled chatbots have facilitated real-time monitoring and analysis of mental health and have made mental health care accessible. With the increasing global prevalence and burden of depression, the time has come for comprehensive mental health policies, increased public awareness and innovative solutions. Predicting mental health using machine learning is an important area of investigation that could bridge the gap between the traditional clinical diagnosis and the advancements in technology. Through the use of data-informed responses and feedback, mental health professionals can develop personalized treatment plans and more effective and efficient mental health interventions that will enhance the lives of individuals with depression.

2.2 Related Works

Rahman et al. represented the calcium dataset in text-based form to accomplish this. (2023) utilized machine-learning to predict depression and anxiety in undergraduate students. Utilizing models such as logistic regression, random forest, neural networks, and KNN; the study classified subjects in minimal, mild, moderate, or severe status and hybrid stacking model with LDA. The hybrid model could predict 99% depression and 97% anxiety and performed better than all classifiers. Their research shows the ensemble increasing classification accuracy. Text-based input, on the other hand, can limit generalizability, and multivariate approaches that incorporate both physiological and behavioral outputs are required to advance the assessment of mental health [1]. Based on NHANES data, Amirhossein et al. (2024) introduced a multimodal machine learning approach in order to predict the severity of depression and to track the individual risk factors. To improve the precision of the predictions, the analysis incorporated demographic, dietary, socioeconomic, lifestyle, medical and clinical information. All the five models which we fitted that are Linear Regression, SVM, XGBoost, Random Forest, LASSO achieved maximum R2 (0. 93) with Random Forest. The paper highlights the importance of multimodal data for improving assessments of mental health. While the limited dataset might affect its generalization [2], this study enhances the prediction of personal depression to contribute to the diagnosis and management of clinical treatments. Trivedi et al. (2022) studied machine learning approaches toward stress, depression, and anxiety, along with the increasing mental health disorder challenges caused by the modern lifestyle stressors. The study identified daily stressors, workplace culture and social elements as major contributors to depression. The best accuracy was achieved by neural networks (97. 2%) on depression prediction among multiple ML models which were compared. Although feature selection and dataset diversity are challenging, the performance of deep learning is still promising in the prediction of mental health. In order to generalize and improve the early detection performance, future research should extend to the multi-modal data [3]. The study by Suhas et al. (2021) applied textual data analysis and machine learning methods to predict early symptoms of depression. The study applied machine learning models to analyze people's answers, if someone reported symptoms like loneliness or disinterest in activities, to assess their likelihood of having depression. RFC model provides the optimal accuracy rate compared with other ML models. An automated system was generated to gather and process text responses for detecting depression. The work highlights the efficiency of text-based methods in early identification but highlights limitations in linguistic variety and generalisation of the dataset. In the future, multimodal input synthesis could be studied to give a more comprehensive picture for depression detection [4]. Mahesh and Amanullah (2023) pitted SVM and KNN algorithms for predicting depression. This study aims at comparing the performances of K Neighbors Classifer and SVM classifier to improve the prediction accuracy. The research used pretest control and statistical significance tests applied on dataset of 20 people and found that KNN is better with an accuracy of 95% compared to the existing system is 84% and the results were "clear" or "acceptable". The statistical analysis (p0. 05) confirmed a significant difference of SVM and KNN, KNN is superior. Their research underlines the utility of distance-based classifiers for predicting depression and the significance of considerably larger data sets for better generalisation [5].

Keya and Han (2022) used five machine learning models (Decision Tree, Random Forest, Multilayer Perceptron (MLP), Support Vector Machine (SVM), and AdaBoost) to investigate the prognosis for depression. They stressed that the best performance by SVM with accuracy of 85%, and minimum false positice, and false negative rate. The research also investigated trends of Depression across different ages and used PCA for the selection of characteristics. Although that ensemble learning could be pursued to further improvement, the performance demonstrates the effectiveness of SVM in depression prediction. It is emphasized in [6] the importance of feature reduction techniques in terms of improving the classification rate. To better diagnose high-risk disease and treat at-20 Shah et al. (2024) applied machine learning models for early prediction of depression. The work employed ten ML models on the Depression Prediction dataset taken from Kaggle and Random Forest provided the best result of 98. 22% accuracy. Moreover, an ensemble of classifiers provided a 0. 98 F1 score, implying its remarkable predictive power. While model generalization and dataset biases remain as challenges, the study suggests the promise of ML in early intervention. Prospective studies may focus on real-time monitoring and multimodal data integration, to improve the identification of depression [7]. Nison et al., based on nonclinical information (such as demographic data, health status, relationship or university life perspective), (2023) proposed a machine learning approach for depression screening in college students. This method is intended to reduce stigma and increase engagement by being free from specific mental health questions, unlike conventional self-report mental health questionnaires. The study accounted for class imbalance by using resampling techniques that improved model performance to an average prediction accuracy of 66%. While this method provides a nonintrusive means identifying depression, further refinements in feature selection and model optimization are required to maximize the accuracy levels for practical use in the clinic [8]. Shinde et al. (2024) used machine learning models to predict depression levels and contributing factors for studying the effects of the COVID19 pandemic on depression. Logistic regression, random forest, SVM, decision tree and KNN to analyze the dataset of 1,504 individuals. According to the study, failure to communicate and sense of guilt were both important predictors of depression. The performance of the classification models to predict depression was shown, underscoring the importance of targeted mental health prevention strategies. Subsequent studies ought to pay attention to monitoring in real time and personalized predictive model to enhance psychological health intervention [9]. Dhawale et al. (2024) adopted on early detection and intervention in their machine learning-based depression prediction research. To model and analyze employee mental health, the study employed the K-Nearest Neighbors (KNN) and Logistic Regression among others for classification. The performance of KNN was better than logistic regression, and the accuracies were 82. 5% and 73. 3%, respectively. The study covered data visualization through the use of heat maps, confusion matrices, pie charts, and bar plots, which demonstrated key trends in mental health. While KNN worked but the work suggests that the performance can be enhanced, in future work, using feature engineering and hybrid model to enhance the performance on prediction and its usability in predicting real-world cases of mental health screening [10].

2.3 Challenges

A major problem of this study was the imbalanced dataset ESPECIALLY IN Depression COLUMN WHERE there were too many NOT depressed (coded as 0) compared to being depressed (coded as 1). The imbalance of the class made it challenging for machine learning algorithms to learn patterns indicative of sadness as most models were biased towards the majority class. The larger false negatives of the models, that is, classifying most individuals who are depressed as being non-depressed, resulted in lower accuracy of depression prediction, which may have been due to significant differences between the two groups according to many demographic and clinical variables. Furthermore, the low recall of the model in identifying the actual depression patients is due to an overfitting to the majority class that impaired the generalization of the model. But maintaining crucial patterns, the set was still difficult to balance. Determining the best method to predict depression and comparing and contrasting different machine learning models was also a significant limitation. Here, the Logistic Regression, SVM, Random Forest, XGBoost, Gradient Boosting and AdaBoost models used in this study have their own pros and cons. Overfitting was an issue especially with complex models such as Gradient Boosting and XGBoost that showed outstanding performance on the training data but failed to generalize the testing data. But the simpler models – the ones like Logistic Regression that had lower recall – better captured the low-level patterns in which depression is manifested. Also, it was hard to tune the hyperparameters, as a lot of iterations and resources were needed just to satisfy that every model was optimally tuned. Since the accuracy, precision, recall, F1 score, and confusion matrices presented different aspects of model performance, it was difficult to make an objective comparison between the models. The aim was to select the model that would not only generate accurate predictions, but one that was also able to accurately detect depression, with no misclassification bias. In the present research, feature selection was another challenge. By means of the recursive feature elimination (RFE) the predictors most relevant to the prediction of the depression were detected. But this was very computationally expensive and quite iterative. While redundant or irrelevant edge-level predictors has to be eliminated for enhanced model performance, it was a tradeoff between the exclusion of valuable predictors. Another problem was the number of features to retain, as too many features might cause the model to lose predictive capacity. For the discrepancies between the weights of the same characteristics in these models, it was difficult to discern what about the characteristics played the most important role in predicting depression. Although there are limitations, feature selection was necessary to reduce overfitting, improve computational performance, and to increase the interpretability of the model. In general, the prediction of depression by machine learning was faced a variety of challenges, such as imbalance data, biased to categorize model and feature selection. The challenges required oversampling techniques, extensive model testing, and advanced feature selection methods. Remove these barriers is critical for developing accurate, reliable and widely applying machine learning models for mental health care. The results of the study, which provides potential strategies for early diagnosis and intervention strategies, add to research surrounding AI-based mental health diagnostics**.**

CHAPTER 3
METHODOLOGY

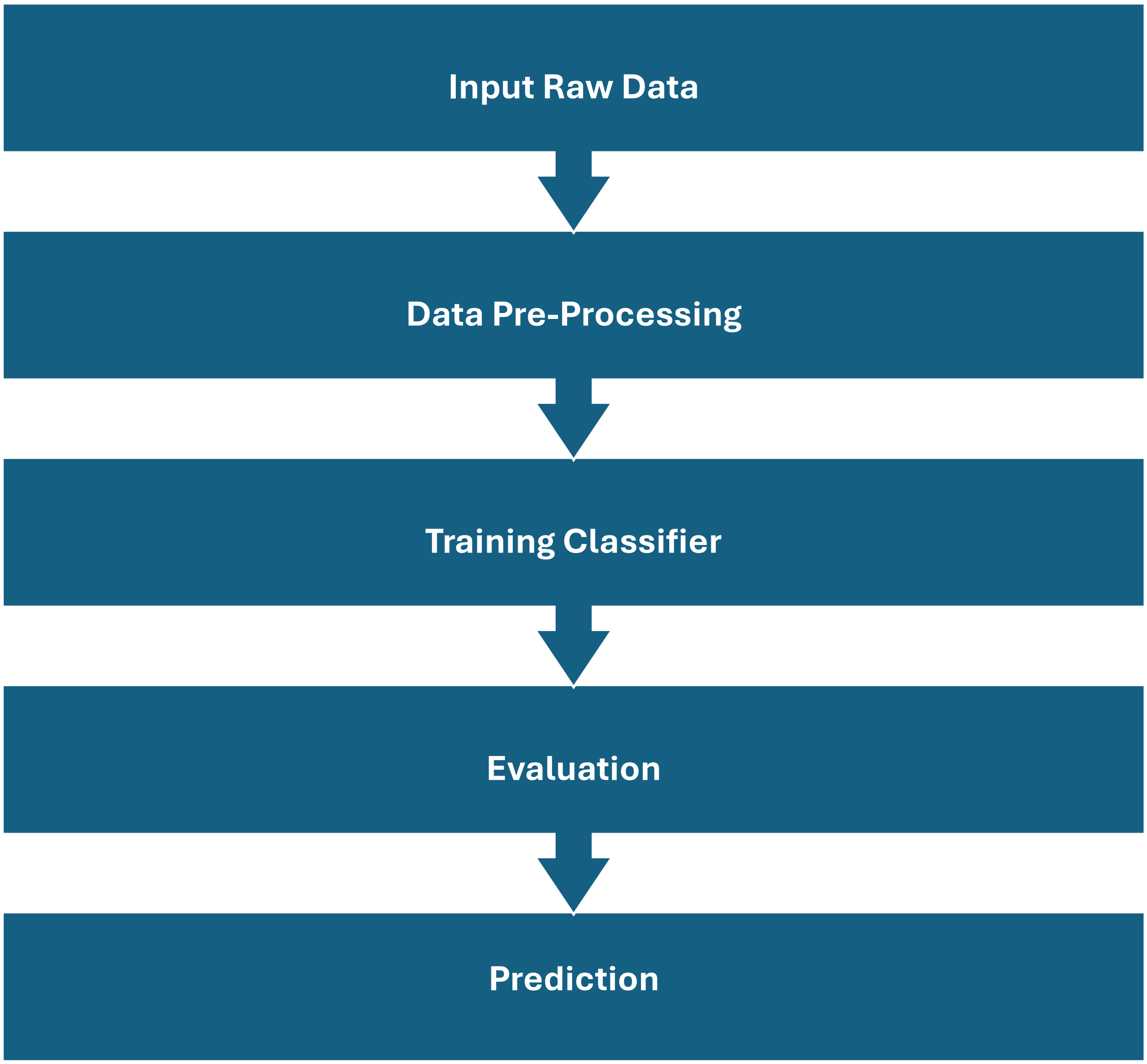

Figure 1.1 Workflow

3.1 Dataset Description

The database In this study, we used a dataset provided in Kaggle's Playground Prediction Competition, season 4, episode 11: exploring mental health data (https://www. kaggle. com/competitions/playgroundseriess4e11). This dataset aims to provide data scientists and researchers an opportunity to explore likely causes of depression using responses to a mental health survey. The dataset could be used to apply machine learning algorithms to predict depression from a range of demographic, psychological, and behavioral variables. This dataset includes many variables which aid us to understand how mental health is related to many individual and environmental factors. Contributing factors include age, gender, financial stress, eating patterns, work and school pressures, the duration of sleep, job satisfaction and having a family history of a mental illness. The target in this dataset is Depression and it is labeled as 0 (Not Depressed) or 1 (Depressed). Because there are far more instances of non-depressed subjects (0) than of depressed ones (1), the data set suffers from a class imbalance problem, as revealed by exploratory data analysis. On the whole, this dataset comes with many features for predicting depression and should provide useful information for the risk factors of mental health problems and how machine learning can be used to enhance early detection and intervention programs. Many features in the dataset measure various aspects of an individual's demographic information, behavior, history of mental health, and work/school related stress that enable the prediction of depression. Each column is a separate feature which can have patterns and correlations to mental health. The id column containing a unique identifier for each individual in the dataset. It doesn't read the Name column while it has individual's name in there, this is usually removed as part of preprocessing for maintaining data privacy and not useful in building predictive model. And if there are any gender trends connected to depression, the Gender column will let you compare men and women. The age in years for each individual which is a key factor in ascertaining whether depression belongs in one or more specific age subgroups was recorded in the Age field. The city column indicates where the person lives, potentially enabling researchers to explore whether there are regional levels of depression. Mental illnesses can be more or less common in a city depending on income levels, cultural differences, or the need to seek mental health services. The Working Professional/Student category classifies individuals based on whether they are students or working professionals due to the fact that these two groups encounter somewhat distinct challenges that would potentially result them in getting depressed. The person's occupation is also defined in the Profession column, which can be used to investigate if different occupations have higher levels of work-related stress and depression. The Academic Pressure column is critical for understanding the impact of academic pressure on mental health, and the mental health of students in particular, and while it sometimes lacks data. Similarly, Work Pressure is indicative of stress associated with work place that is a known contributing factor to anxiety and depressive disorders. An interesting finding in the case of students is the CGPA: it is in fact a measure of academic success and may be associated with mental health issues in underachievers, i.e., stress/burnout or low self-esteem. Sleep Duration: the hours spent sleeping per night is entered in the Sleep Duration column, categorized into ranges such as Less than 5 hours or Over 8 hours. Job Satisfaction indicates the happiness among professionals at work, and (the) Study Satisfaction for student in their academic development. When either of these out- comes are disappointing, this can be an important symptom of depression. This is an important one as sleep deprivation is a known cause of poor mental health. The Dietary Habits: This column categorizes the individual dietary pattern in Healthy, Unhealthy and Moderate. Nutrition has an effect on mental health, and junk food in particular is increasingly being linked to depression and the like, so when your diet sucks, so does your mood. The Degree column indicates a person's highest degree, information that can provide insight into whether the level of education relates to mental health issues. Have you ever had suicidal thoughts? question in the poll is whether someone has had suicidal thoughts, a key indicator of severe depression. The number of hours worked or studied per day was noted in the Work/Study Hours.06 This might be potential mental health stress when it is over. Financial Stress measures people's lack of liquidity, a major cause of depression, especially in students and younger employees. And whether someone is genetically predisposed to having problems with mental health, which is a big factor in depression, is exposed through the Family History of Mental Illness section. The column Depression indicates whether a person is depressed (1) or not (0) and is the target variable. This binary distinction will serve as a model to the prediction model in this study. This dataset is valuable for employing machine learning in examining and predicting depression, owing to the comprehensive information regarding variables that impact mental health. The aim of this study is to develop an effective model for the early detection of depression and improving mental health support systems by investigating the relationships among such issues..

Table 3.1 Dataset

Gender	Age	City	Working Professional or Student	Profession	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness	Depression
Female	49.0	Ludhiana	Working Professional	Chef	NaN	5.0	NaN	NaN	2.0	More than 8 hours	Healthy	BHM	No	1.0	2.0	No	0
Male	26.0	Varanasi	Working Professional	Teacher	NaN	4.0	NaN	NaN	3.0	Less than 5 hours	Unhealthy	LLB	Yes	7.0	3.0	No	1
Male	33.0	Visakhapatnam	Student	NaN	5.0	NaN	8.97	2.0	NaN	5-6 hours	Healthy	B.Pharm	Yes	3.0	1.0	No	1
Male	22.0	Mumbai	Working Professional	Teacher	NaN	5.0	NaN	NaN	1.0	Less than 5 hours	Moderate	BBA	Yes	10.0	1.0	Yes	1
Female	30.0	Kanpur	Working Professional	Business Analyst	NaN	1.0	NaN	NaN	1.0	5-6 hours	Unhealthy	BBA	Yes	9.0	4.0	Yes	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Female	18.0	Ahmedabad	Working Professional	NaN	NaN	5.0	NaN	NaN	4.0	5-6 hours	Unhealthy	Class 12	No	2.0	4.0	Yes	1
Female	41.0	Hyderabad	Working Professional	Content Writer	NaN	5.0	NaN	NaN	4.0	7-8 hours	Moderate	B.Tech	Yes	6.0	5.0	Yes	0
Female	24.0	Kolkata	Working Professional	Marketing Manager	NaN	3.0	NaN	NaN	1.0	More than 8 hours	Moderate	B.Com	No	4.0	4.0	No	0
Female	49.0	Srinagar	Working Professional	Plumber	NaN	5.0	NaN	NaN	2.0	5-6 hours	Moderate	ME	Yes	10.0	1.0	No	0
Male	27.0	Patna	Student	NaN	4.0	NaN	9.24	1.0	NaN	Less than 5 hours	Healthy	BCA	Yes	2.0	3.0	Yes	1

3.2 Exploratory Data Analysis (EDA)

3.2.1 Missing value

Missing values in several columns make data preparation and model accuracy a struggle. The dataset contains a large amount of missing data for key variables such as Academic Pressure (112,803 NaNs), Study Satisfaction (112,803 NaNs), Work Pressure (27,918 NaNs), and Job Satisfaction (27,910 NaNs), implying many survey respondents did not provide data for these variables. Profession Professional 36,630 missing You could assume from the value distribution of the Profession column, some people might skip to answer this question or maybe it does not matter for all if we all assumed to have a profession thus is an unnecessary c Feature and should be dropped. Furthermore, some categorical variables, like Dietary Habits, Degree, and Family History of Mental Illness, are very low on missing values and can be handled with mode imputation, that is, replacing missing values with the category that occur most frequently. The first two columns, Work/Study Hours and Financial Stress, also contain very few missing values, for which we can use median imputation.

3.2.2 Statistical Summary

The dataset we describe in this work has 140,700 records of features such as the age, the academic pressure, the work pressure, the {CGPA}, the study satisfaction, the job satisfaction, the work/study hours, the financial stress, and the depression s tatus. The age of the patients ranges from 18 to 60 years, with a mean of 40. 39 years, showing a spread in the age of participants including students and employed people. The average scores of work pressure and academic pressure are both 3.14 and 2.99, respectively (1 – 5 scale points); the work pressure and academic pressure …are moderate. CGPA CGPA (grades) is an indicator of the different academic performances of students and ranges between 5. 03 to 10. 00, with an average of 7. 56. Mean values for respondents rated their satisfaction overall in professional/academic life were moderate overall (2. 94 and 2. 97 for study and job satisfaction (1 – 5 scale).

Table 3.2: Statistical Summary

	id	Age	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Work/Study Hours	Financial Stress	Depression
count	140700.000000	140700.000000	27897.000000	112782.000000	27898.000000	27897.000000	112790.000000	140700.000000	140696.000000	140700.000000
mean	70349.500000	40.388621	3.142273	2.998998	7.658636	2.944940	2.974404	6.252679	2.988983	0.181713
std	40616.735775	12.384099	1.380457	1.405771	1.464466	1.360197	1.416078	3.853615	1.413633	0.385609
min	0.000000	18.000000	1.000000	1.000000	5.030000	1.000000	1.000000	0.000000	1.000000	0.000000
25%	35174.750000	29.000000	2.000000	2.000000	6.290000	2.000000	2.000000	3.000000	2.000000	0.000000
50%	70349.500000	42.000000	3.000000	3.000000	7.770000	3.000000	3.000000	6.000000	3.000000	0.000000
75%	105524.250000	51.000000	4.000000	4.000000	8.920000	4.000000	4.000000	10.000000	4.000000	0.000000
max	140699.000000	60.000000	5.000000	5.000000	10.000000	5.000000	5.000000	12.000000	5.000000	1.000000

3.2.3 Categorical column visualization

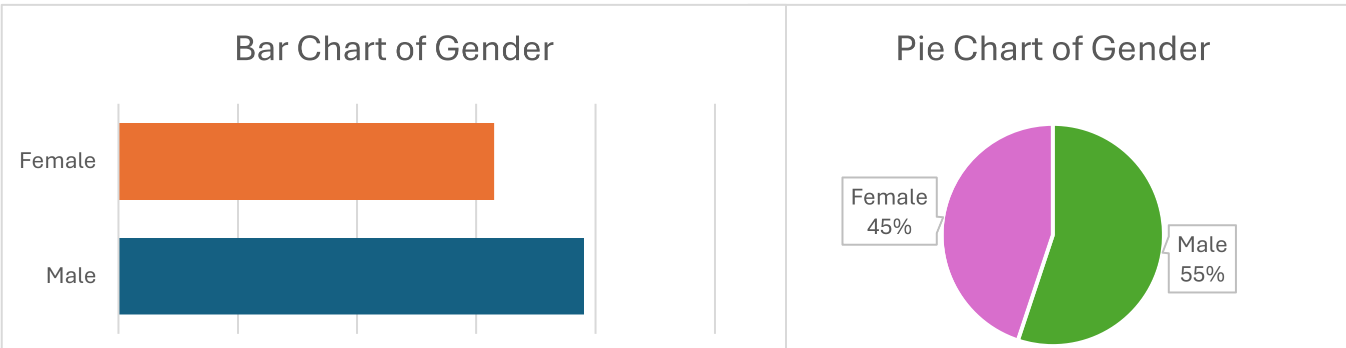

Figure 3.1: Visualizations of Gender

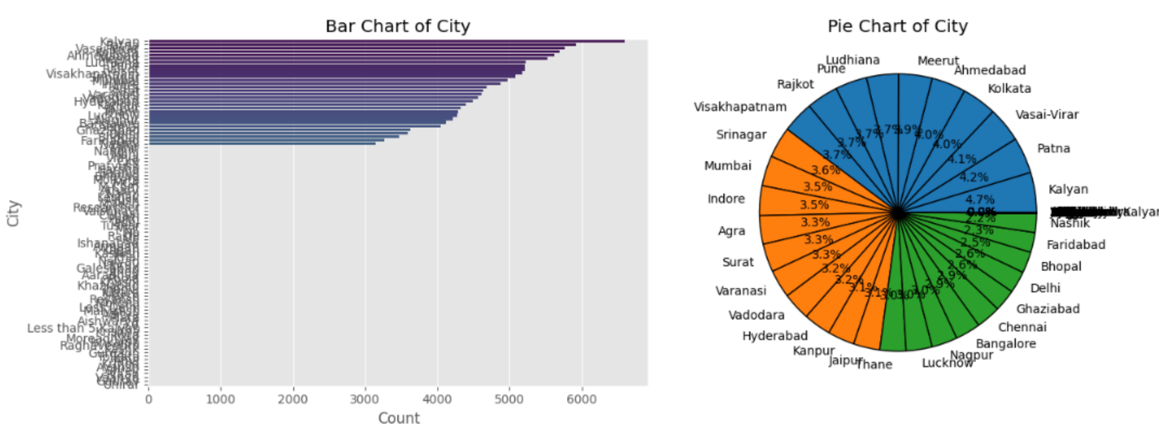

Figure 3.2 Visualizations of City

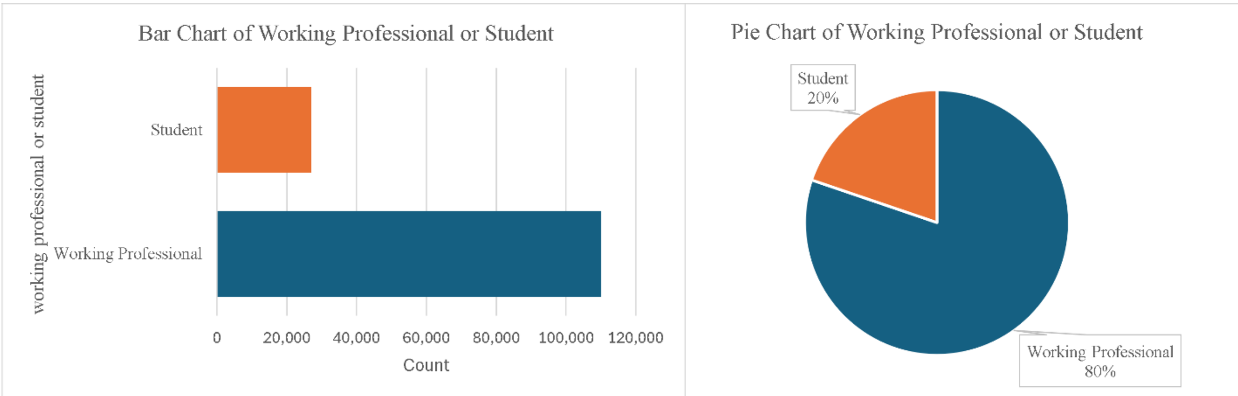

Figure 3.3: Visualizations of Working Professional Or Student

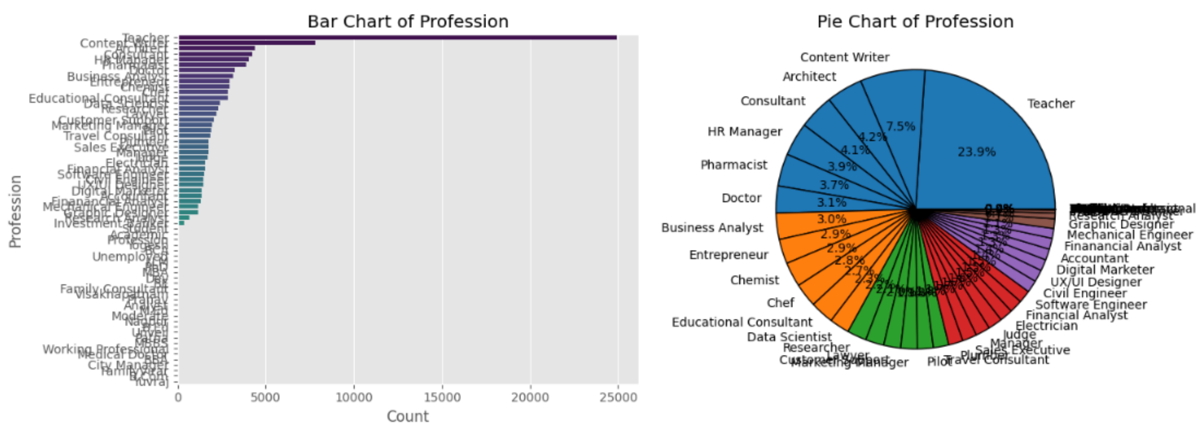

Figure 3.4: Visualizations of Profession

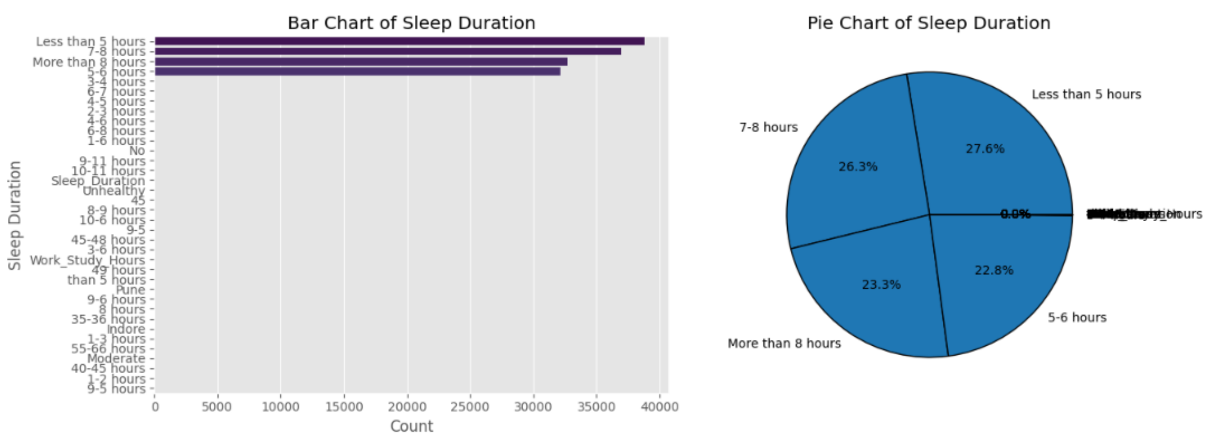

Figure 3.5: Visualization of Sleep Duration

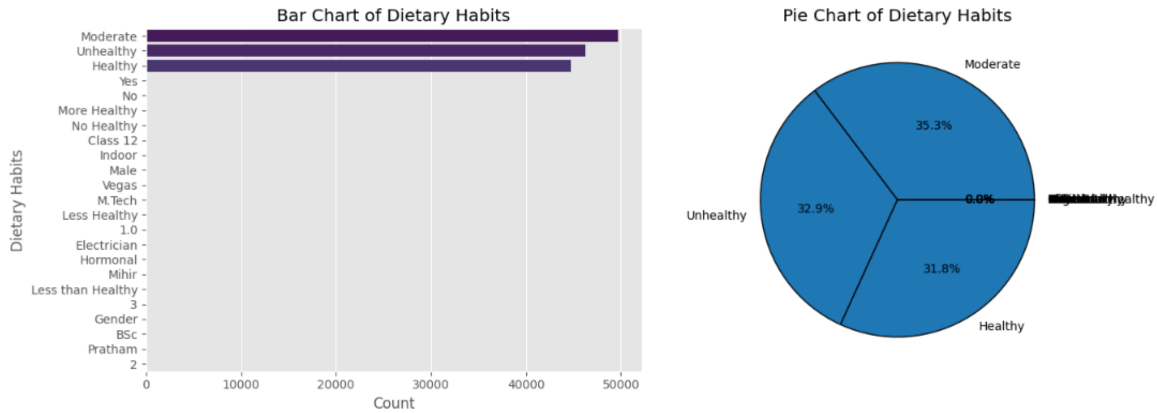

Figure 3.6: Visualizations of Dietary Habits

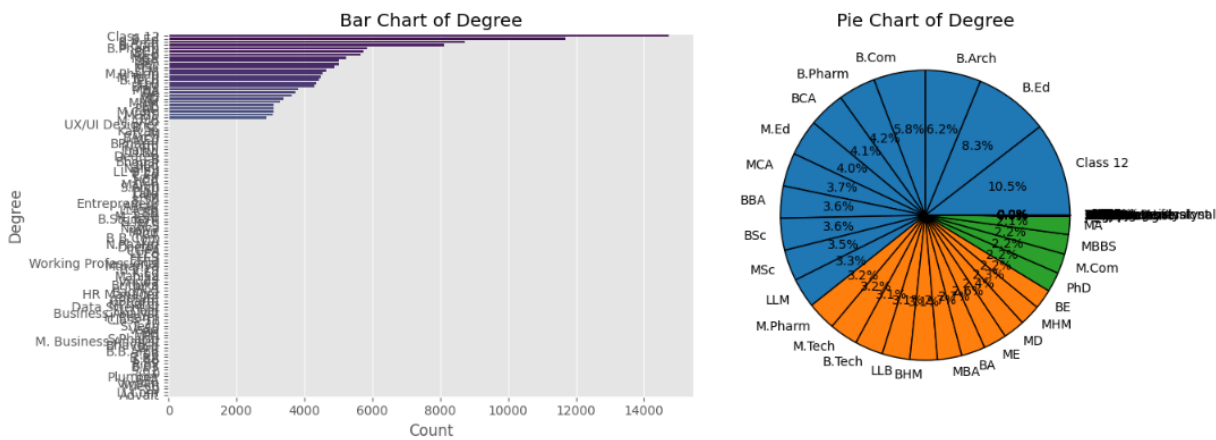

Figure 3.7: Visualizations of Degree

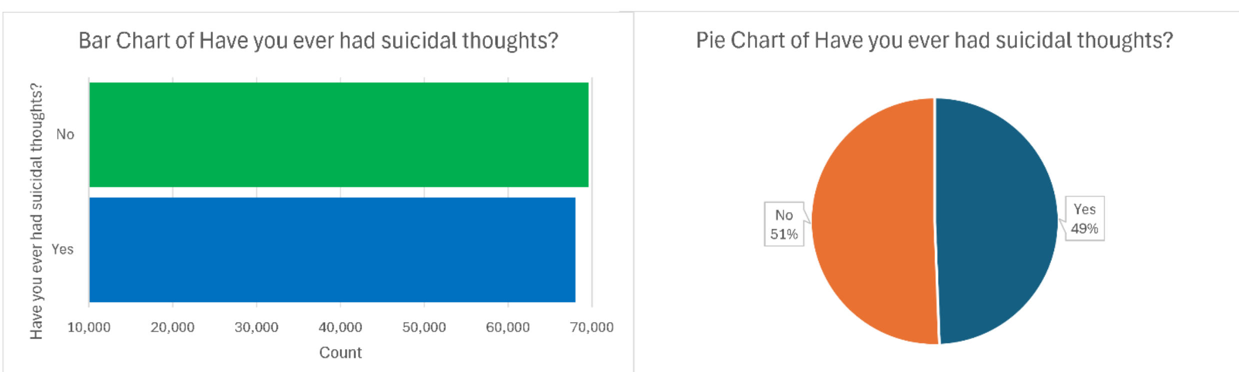

Figure 3.8: Visualizations of Have you ever had suicidal thoughts?

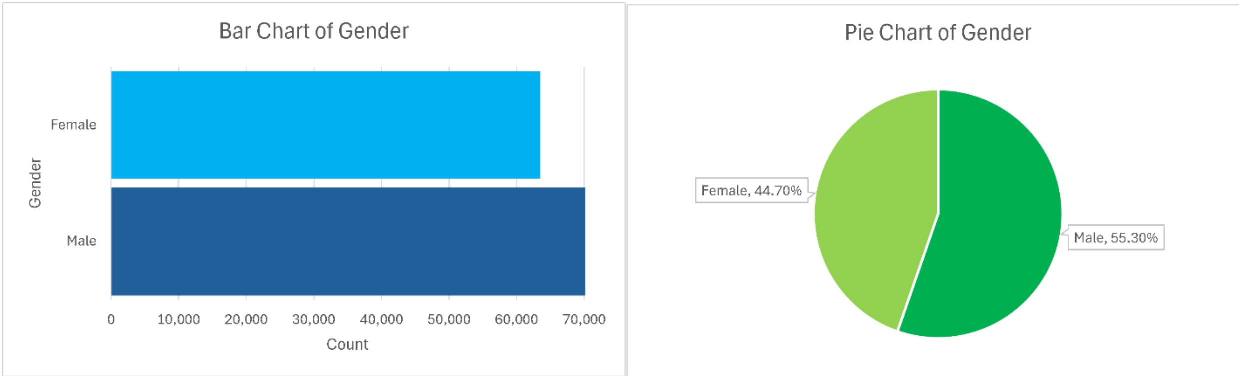

Figure 3.9: Visualizations of Gender

3.2.4 Numerical Column Visualization

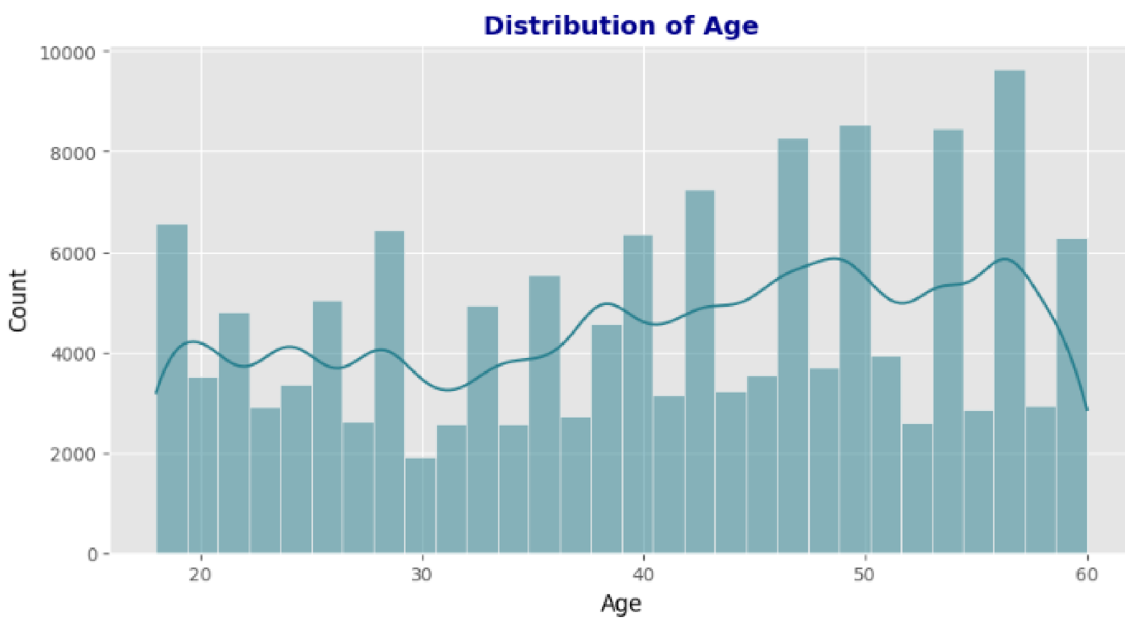

Figure 3.10: Visualizations of Distribution of Age

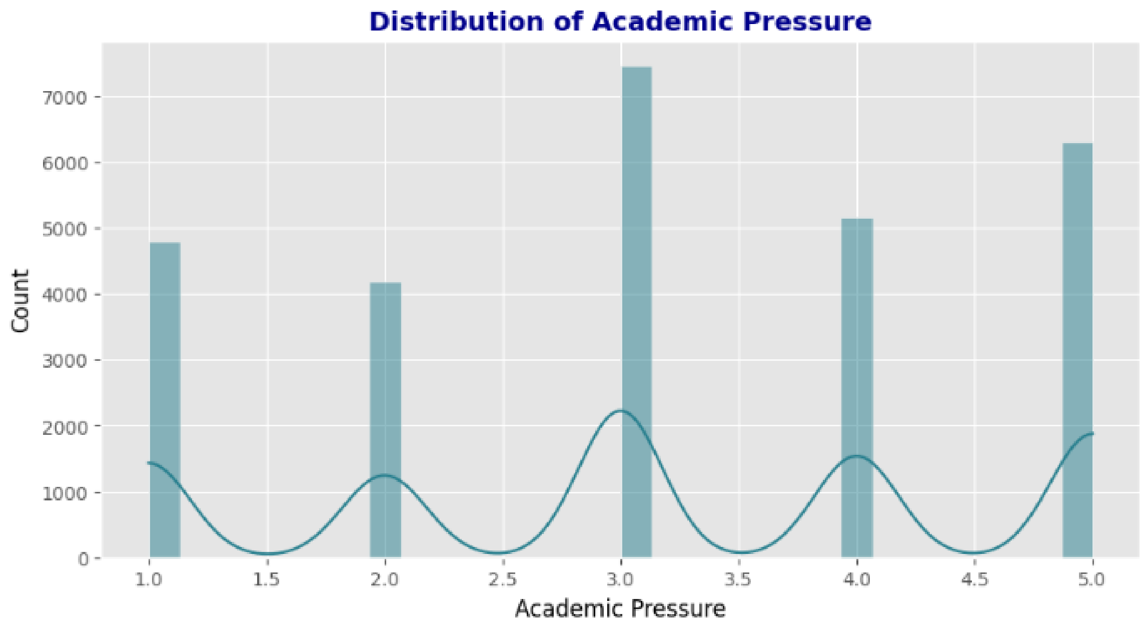

Figure 3.11: Visualizations of Distribution of Academic Pressure

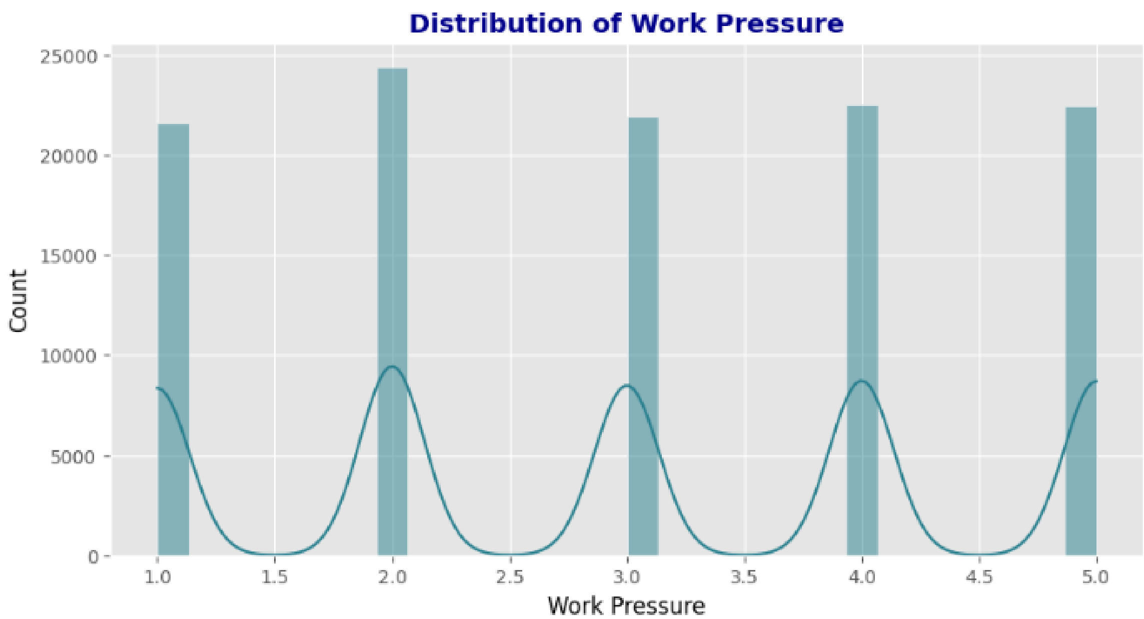

Figure 3.12: Visualizations of Distribution of Work Pressure

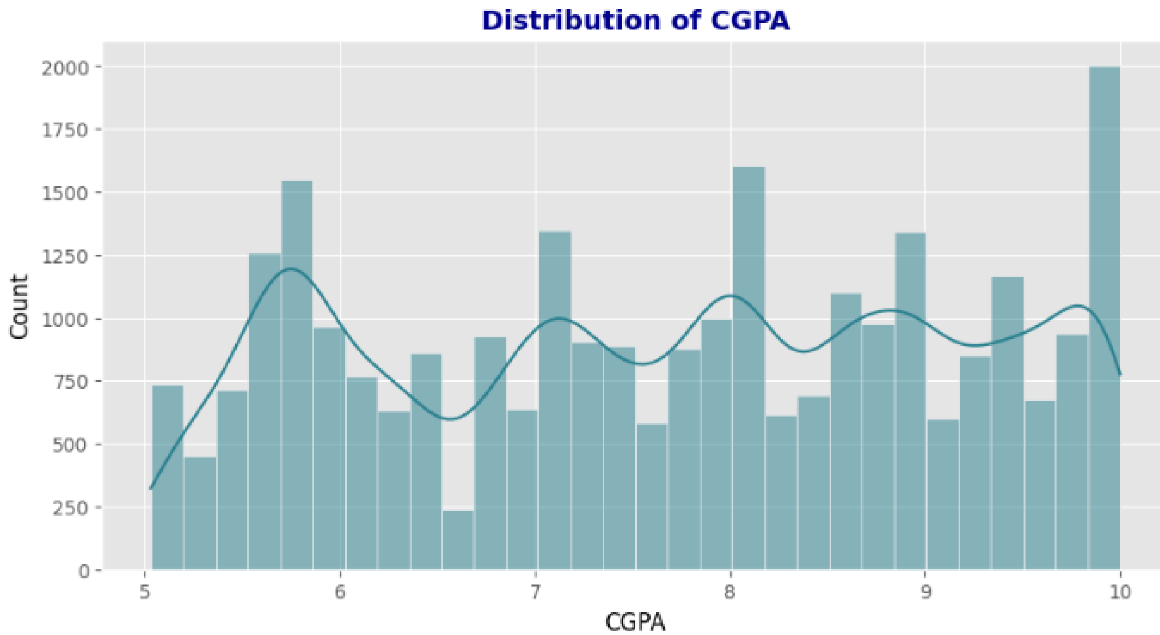

Figure 3.13: Visualizations of distribution of CGPA

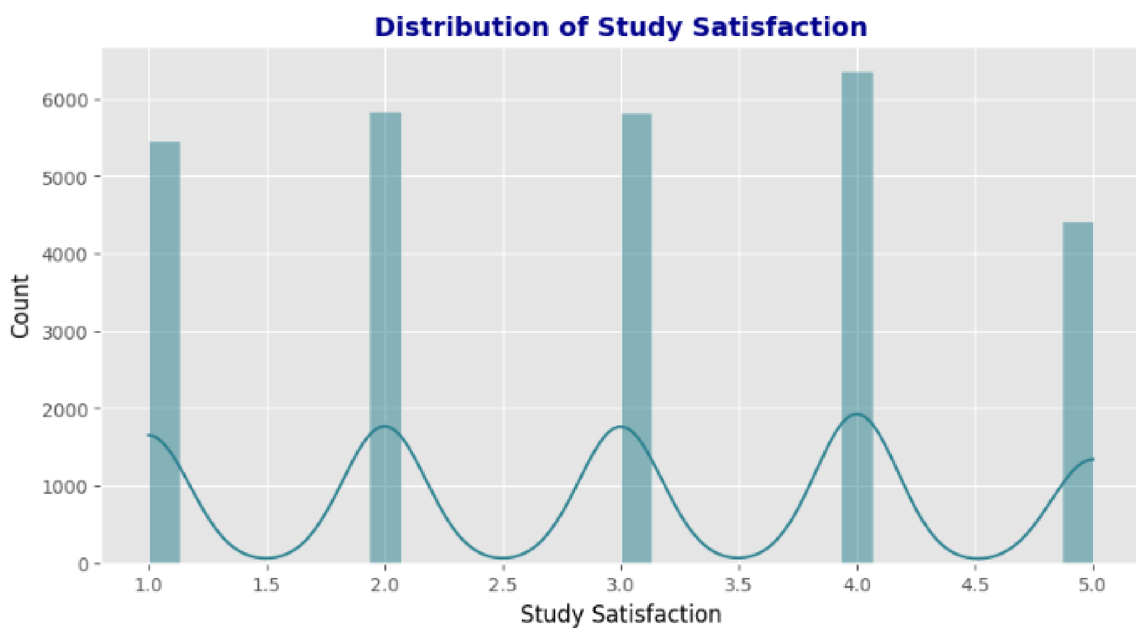

Figure 3.14: Visualizations of Distribution of Study Satisfaction

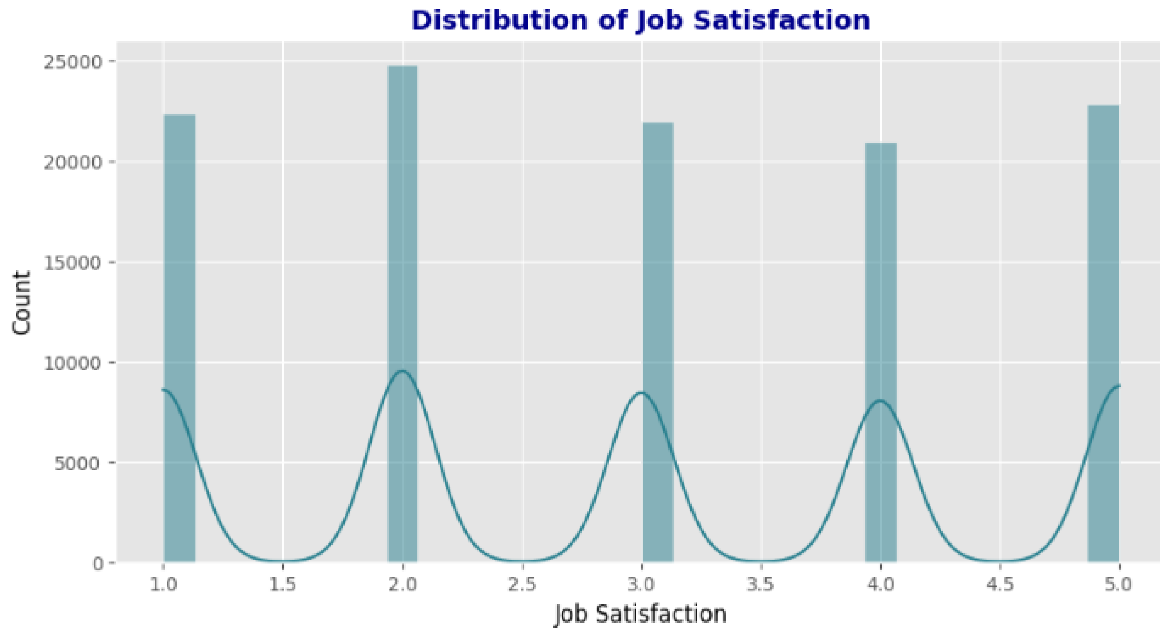

Figure 3.15: Visualizations of Distribution of Job Satisfaction

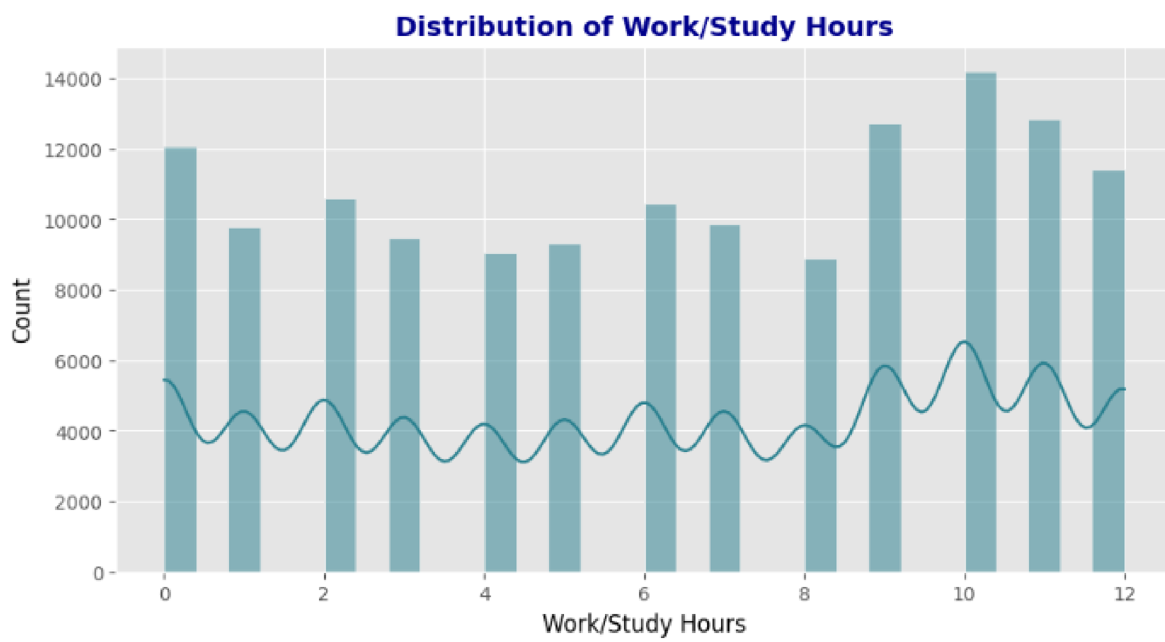

Figure 3.16: Visualizations of Distribution of Work/Study Hours

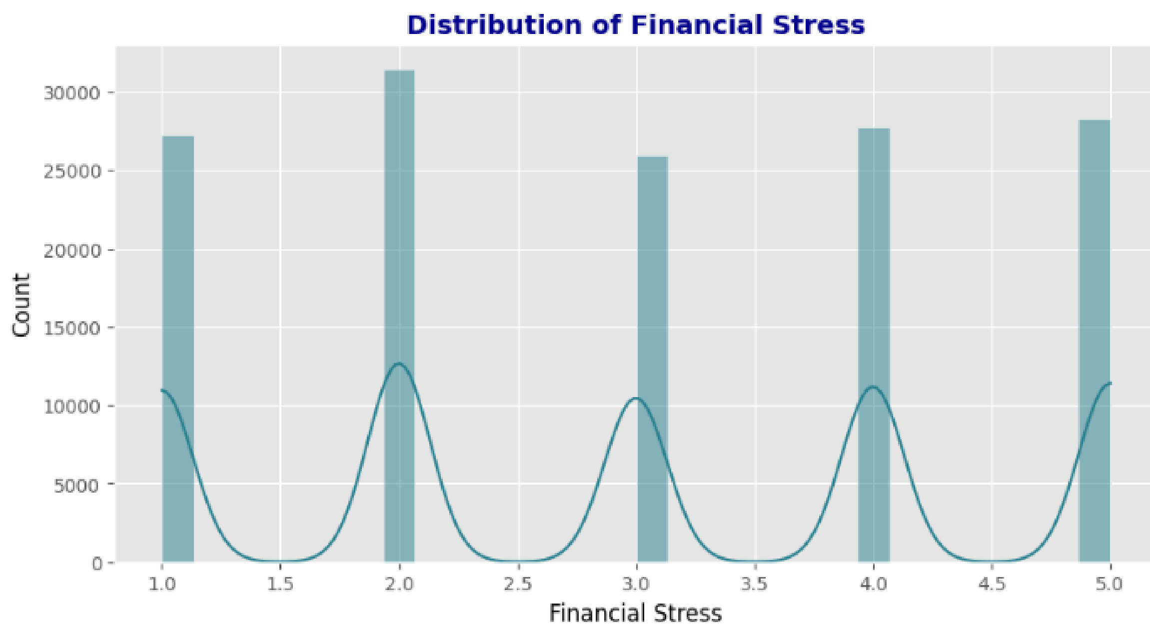

Figure 3.17: Visualizations of Distribution of Financial Stress

3.3 Data Preprocessing

In this paper, we focus on examining data preprocessing at the data level to enhance the performance (i.e., the accuracy of the models and the reliability of the predic- tions) of the depression predictions built using machine models. The model's precision might be affected by outliers in the data. Statistical approaches like IQR, and zsore analysis etc were used to handle the outliers, which can bias the model's predictions. In the data, missing values are another serious issue, as some columns (e.g., Academic Pressure and Study Satisfaction) even have high missing rate. To remove bias and allow better interpretation of the model, columns with too many missing values were dropped. For characteristics with a reasonable amount of missing data the mean or median imputation approach was applied. For factors like Work Pressure, CGPA, Financial Stress the numerical mean was used as imputation and for factors like Dietary Habits and Family History of Mental illness the mean was used as median to make sure that the consistency was kept same.

Moreover, feature scaling was performed to normalize numerical values to further improve the ability of models to process data. This preprocessing pipeline was employed for cleaning, balancing and structuring the dataset for optimizing model training and performance evaluation.

3.4 Feature Engineering

Feature engineering, a crucial step to improve the predicted accuracy and performance of machine learning, is the selection of the most important attributes that contribute to depression classification. Recursive Feature Elimination (RFE) which is a popular feature selection approach is employed to select the key drivers of depression in this study. Since it gradually eliminates unimportant features unless the optimal subset is found, RFE is an iterative process and model automatically concentrates on important features of the predictor, and avoids undesirable information. In this study, the Random Forest Classifier was selected as the base model of the RFE whose selected features were given in order of their importance to the classification.

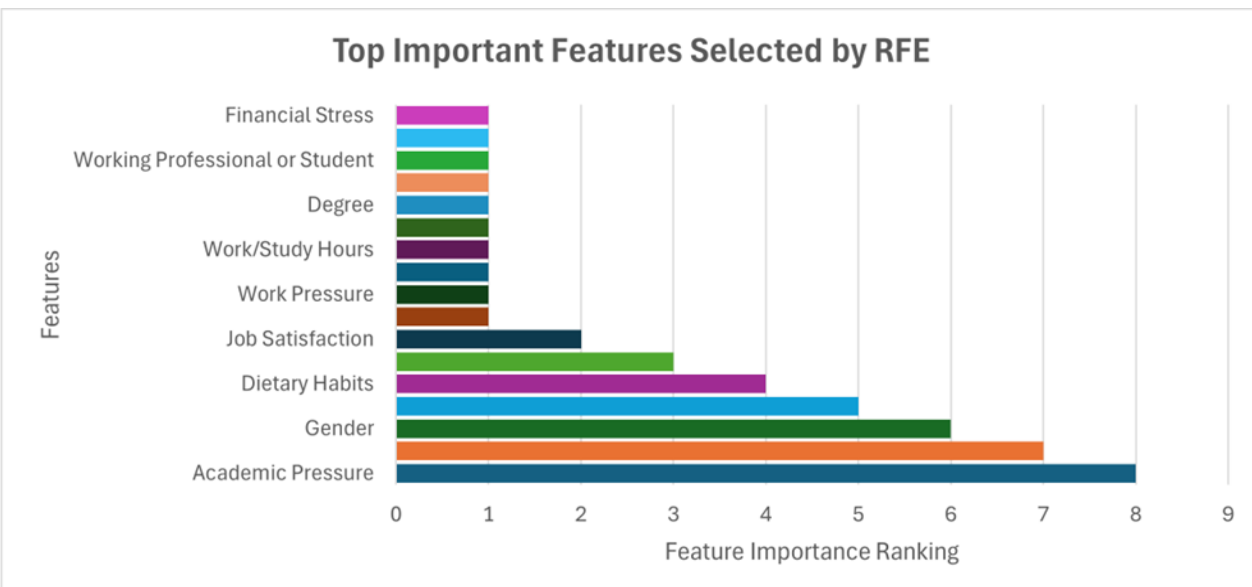

Figure 3.18: Top important features selected by RFE

3.5 Machine Learning Models Applied

3.5.1 XGBoost (Extreme Gradient Boosting)

The XGBoost gradient boosting implementation known as XGBoost is a well-tuned implementation and is extremely fast and powerful. It uses an ensemble learning process that involves creating a number of decision trees, each of which corrects the flaws of the ones before it. The gradient descent is applied to minimize the loss function; this is the reason that the method is highly successful for the high dimensional and unbalanced datasets. To control over-fitting XGBoost has pruning in forms of the maximum_depth of the tree and the min_child_weight for the minimum number of instances in a leaf; the L1 and L2 regularization is used in the objective function which defines the loss that needs to be minimized when growing trees. Furthermore, due to its ability to do parallel computations it is much faster than conventional gradient boosting methods. For Media Depression is one of the common clinical indicators in predictive modeling, and XGBoost is widely used due to its scalability, high classification accuracy, and longevities.

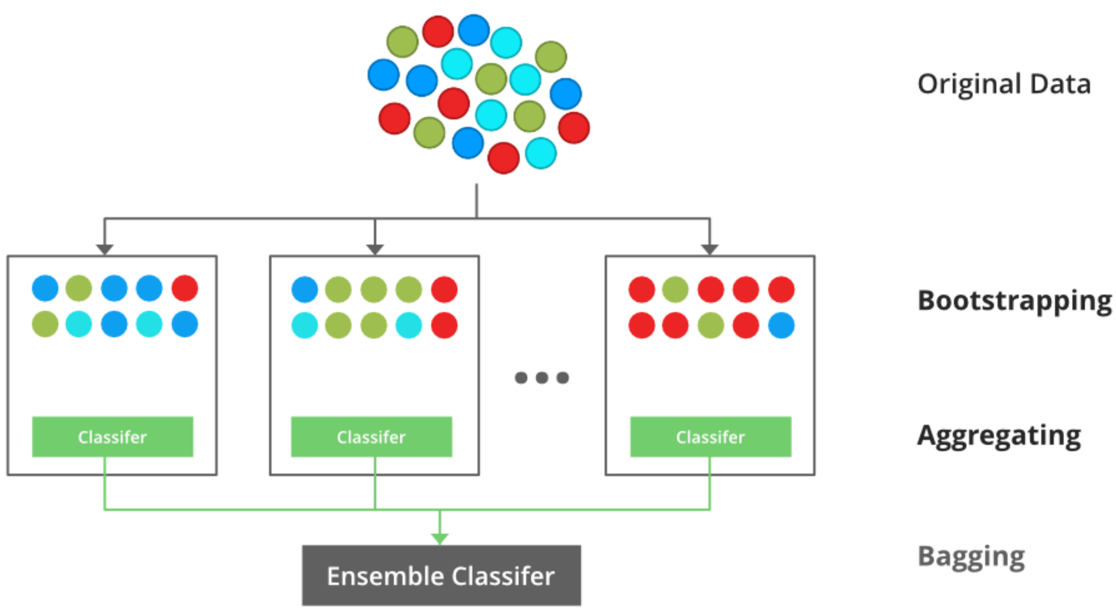

Figure 3.19: XGBoost (Extreme Gradient Boosting)

3.5.2 Gradient Boosting Classifier (GBC)

The Gradient Boosting Classifier (GBC), that is an ensemble learning method where a series of weak learners, often decision trees, are added in continuity where each tree is fixing the errors of its predecessor. In contrast to the classical decision tree models, GBC uses the gradient descent to repeatedly improve the prediction in order to minimize a loss function. The advantage of GBC lies in it capability for identifying complex nonlinear pattern correlations and also its flexibility to work with different forms of data distributions. But when not fine-tuned with regularization such as subsampling, learning rate decay, and tree depth limitation, GBC is both expensive to train and overfitting-natured. Through exposing complex associations between the lifestyle factors and mental health indicators, the GBC contributes to the prediction of depression.

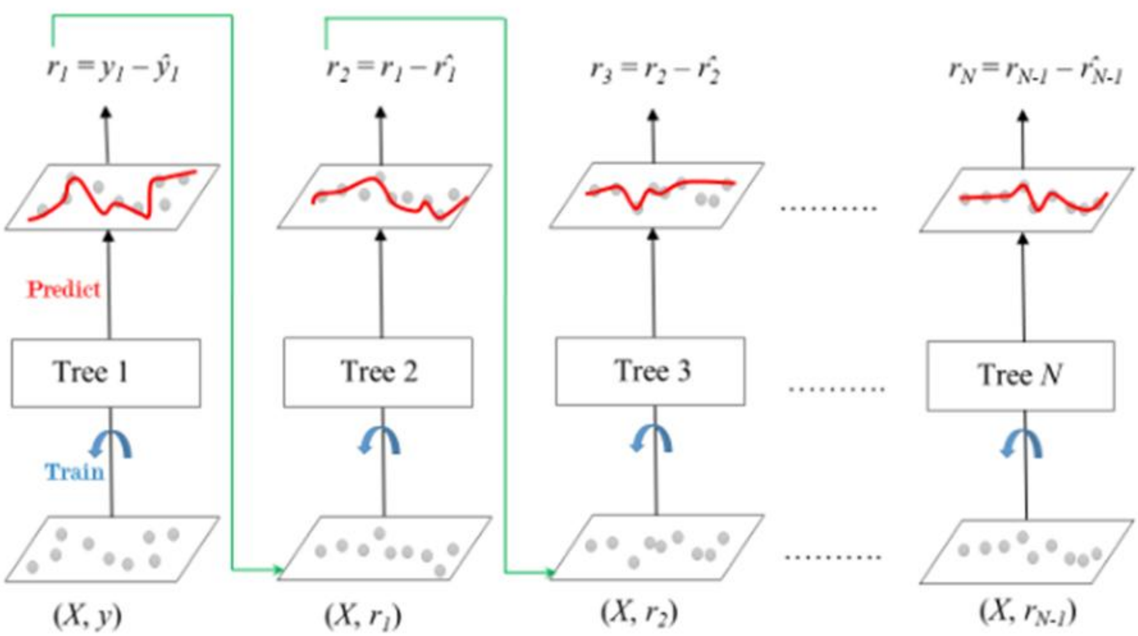

Figure 3.20: Gradient Boosting Classifier (GBC)

3.5.3 AdaBoost (Adaptive Boosting) Classifier

AdaBoost, short for Adaptive Boosting, is an ensemble learning technique that combines the predictions of several weak classifiers, most often 1-level decision trees, to produce highly accurate predictions. The most basic premise of AdaBoost is that it gradually shifts the emphasis to cover the incorrectly classified samples with different weights in each stage. This model incrementally updates weights to optimize discrimination, while focusing on the most challenging examples. AdaBoost, unlike XGBoost and Gradient Boosting does not utilise gradient descent to update the sample importance, but rather it also uses misclassification rates. It can be especially beneficial for binary classification problems where traditional methods may fail to detect subtle patterns in features (e.g., depression detection). AdaBoost itself might lose stability due to its sensitivity to outliers and noisy data.

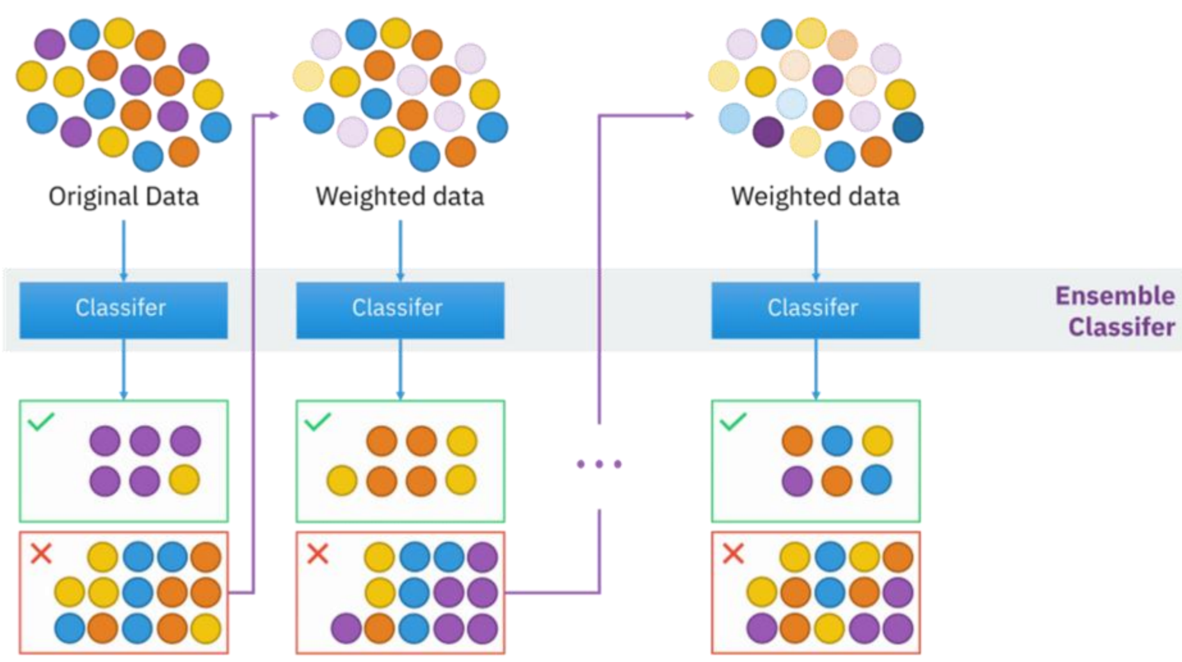

Figure 3.21: AdaBoost (Adaptive Boosting) Classifier

3.5.4 Logistic Regression

The main application for the supervised learning technique logistic regression. The model is based on the logistic (sigmoid) function that maps the linear combination of the input variables into a probability value in the range [0,1]. Equation (3) is used by the model to forecast whether an event will belong to a certain category:

Equation 1 Logistic Regression

$$\phi(z) = \frac{1}{\{1 + e^{\{-z\}}\}} \tag{1}$$

with the βbeta β values being the learned coefficients from the data. Linearly separable data is a good fit for logistic regression and it is also really intuitive. This work would treat it as the standard model to be compared to more sophisticated ensemble methods such as XGBoost, GBC and AdaBoost. However, logistic regression itself is perhaps not sufficient, given depression is a complex disease with nonlinear relationships among indicators. Logistic regression performs better when used with feature selection methods such as RFE (Recursive Feature Elimination)..

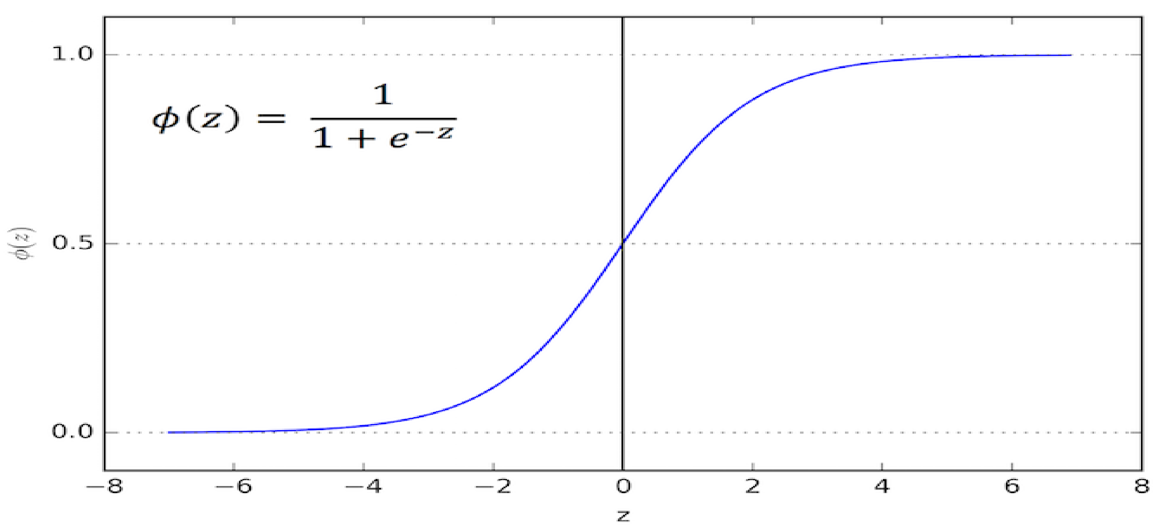

Figure 3.22: Logistic Regression

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Model Performance Comparison

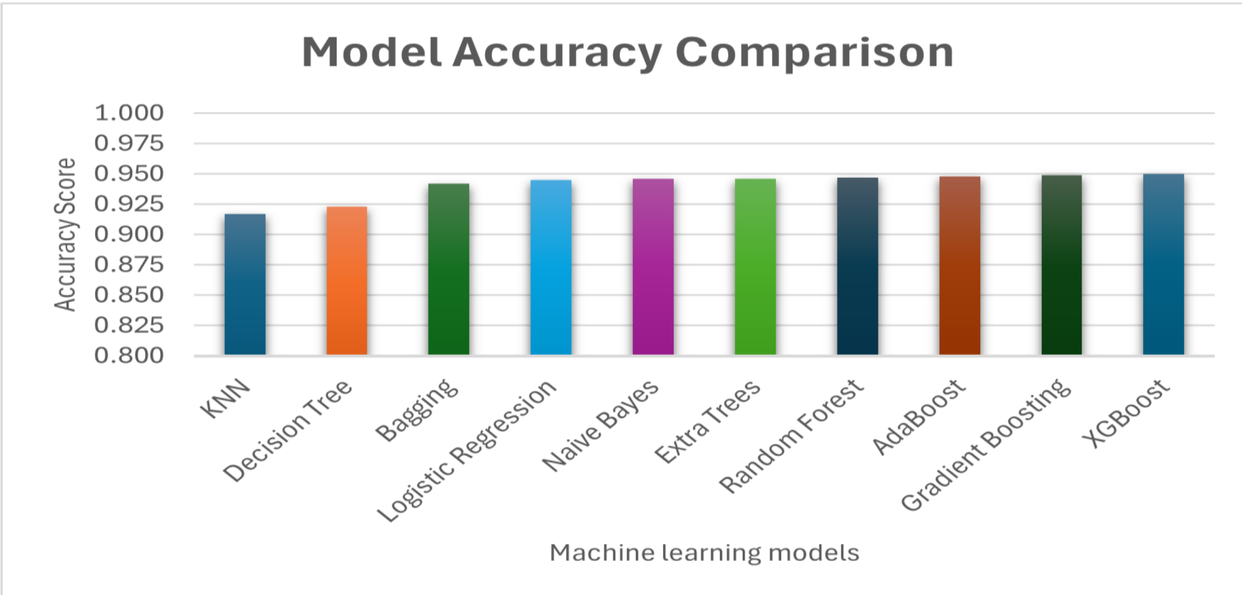

Figure 4.1: Model Accuracy Comparison

The bar graph provides accuracy scores (expressed in %) of the machine learning algorithms in depression prediction. The y-axis shows accuracies for different models of machine learning and the x-axis shows the models. Based on the chart, AdaBoost, Gradient Boosting and XGBoost achieved the highest accuracy, which indicated that they were the best at recognizing the depressed instances. Other ensemble methods such as Naïve Bayes, Extra Trees, and Random Forest are beneficial as well since they give good results, demonstrating strong forecasting power. Additionally, the lower accuracy of LR, Bagging, Decision Tree and KNN models suggests that they may not be handling the variances in depression data as efficiently. Being sensitive to class imbalance and high dimensional datasets, KNN presumably had the least accuracy. As a general trend, ensemble methods outperformed conventional models, thereby stressing the usefulness of boosting and bagging in improving classification accuracy..

4.2 Classification Reports and Confusion Matrices

Classification Report – XGBoost:

Table 4.1: Classification Report – XGBoost

	Precision	***Recall***	***F1-score***	***Support***
0	0.97	0.98	0.97	21310
1	0.80	0.74	0.77	2744
Accuracy			0.95	24054
Macro avg	0.88	0.86	0.87	24054
Weighted avg	0.95	0.95	0.95	24054

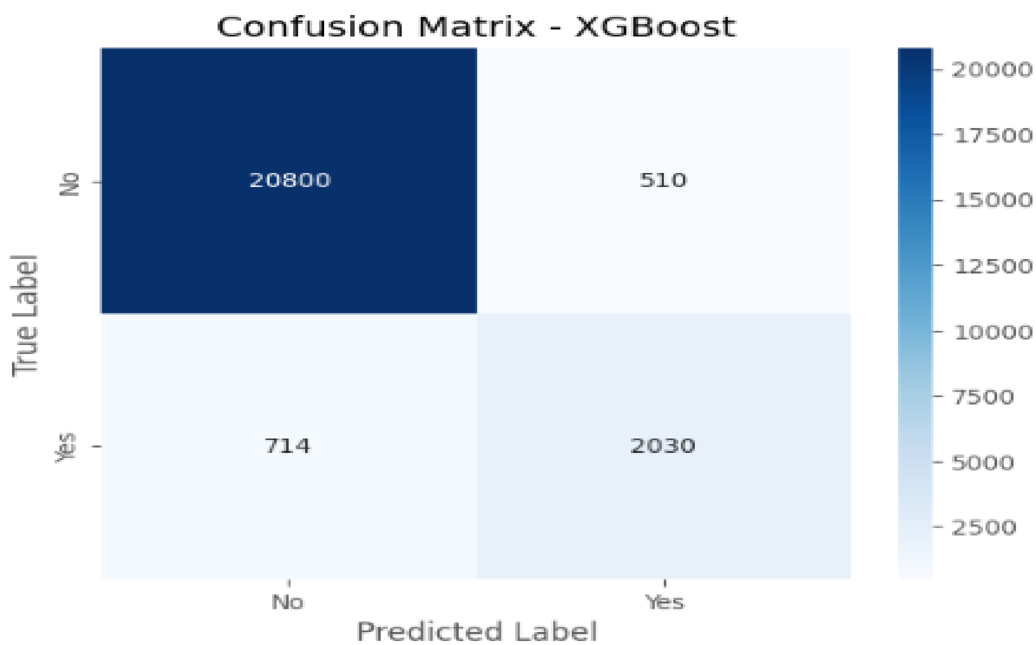

Figure 4.2: Confusion Matrix - XGBoost

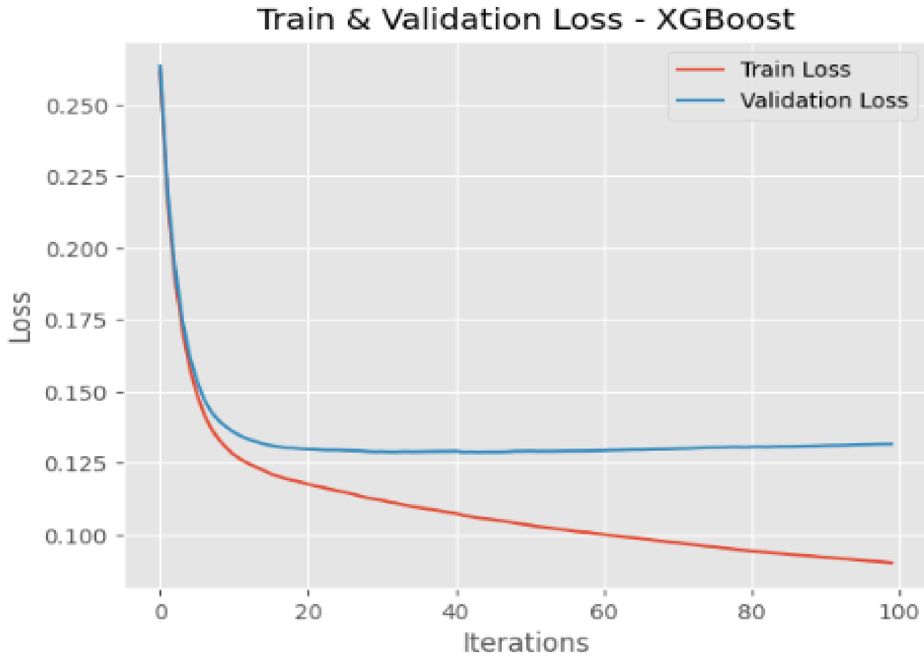

Figure 4.3: Train & Validation Loss – XGBoost

4.3 Impact of RFE on Model Performance

The machine learning models applied to depression prediction in this paper were significantly enhanced in both efficiency and interpretability as a consequence of RFE application. The performance of the model-trained with RFE was similar to the un-selected model (95%), indicating that the removal of features did not cause any negative impact on performance. This demonstrates the effectiveness of RFE in reducing dataset complexity while retaining model performance. RFE assisted to reduce dimensionality by eliminating the less important features in the datasets, thus improving the model efficiency and computing time. It is also beneficial, particularly when working with large datasets, since too many features can result in overfitting, longer training times, and additional noise in the model. The successful application of RFE in the presented work shows that a smaller, more relev- ant subset of features could still yield the same level of accuracy as the whole dataset, leading to a more interpretable and efficient model without compromising performance. Overall, RFE was an effective feature selection method, enabling the construction of a depression prediction model that was more straightforward, faster, and more applicable to general cases. With such reduction in dimension i.e., size of datset and removal of redundant variables, this technique could be applied to other mental health studies to gain efficiency of the model.

4.4 Discussion of Findings

The results of this study demonstrate that prediction of depression is very good by using a number of behavioral, demographic and psychological factors, where machine learning models performs very well. Various models were compared such as theXGBoost, Gradient Boosting, AdaBoost, Random Forest, and Logistic Regression, and in all cases, the ensemble models yielded better performance than traditional classifiers. Among them, Gradient Boosting and XGBoost had the best accuracy, which evidenced the capability for handling complex dataset correlations. The relatively low performance of simple methods such as K Nearest Neighbors (KNN) and Decision Tree suggests that ensemble methods are more appropriate for the task of depression prediction. The proper utilization of RFE was a significant finding of this study. Dropping irrelevant attributes did not change the categorization accuracy because the model has never obtained less than 95% when using or not using RFE. This shows good RFE is at reducing the dimensions in a data set, but still produce good predicative results, since it is much cheaper than an exponentially large set. Eliminating irrelevant features also reduced the likelihood of overfitting, which ensures that the model generalize well on new data. Another observation was that the number of people classified as “Not Depressed (0)” was significantly higher than the number of people classified as “Depressed (1)”, which is referred to as class imbalance problem. The high-class imbalance becomes a challenge for model training because models tend to be biased towards predicting the majority class. The examination of key parameters also provided useful insights about the factors influencing depression. Stress at work, worrying about money, too little sleep, service in the Armed Forces, satisfaction with time at work, and family history of mental illness were all useful predictors of depression. And let’s face it it’s common sense, too, with psychological research telling us that there is very likely to be a powerful interplay between work-related stress, a lack of cash and genetic vulnerability in the development of mental ill health. In this sense, machine learning models have potential for real-world use in early intervention and mental health screening programme for identification of these patterns. In general, this work shows that machine learning can be a valuable tool for predicting mental health outcomes, providing a fair, quick, and scalable mechanism for identifying candidates who might become depressed. Results show the importance of feature selection, ensemble models and class balancing methods for enhancing the prediction accuracy. This could be extended in the future by adding physiological markers, sentiment analysis of data from social media, or in-the-moment monitoring data, to improve detection of depression even more.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Machine learning (ML) plays a significant role in the anticipation of depression, which can result in beneficial mental health care with a positive social impact. One of the leading global health problems, depression, has affected more than 280 million people all over the world. However, it is often undiagnosed or untreated because of stigma, lack of knowledge or access to mental health specialists. ML may be able to bridge this disconnect, by creating an early detection system that provides proactive, scalable, low-cost mental health treatment.By embedding depression prediction models into digital platforms, such as telehealth systems, chatbots, or mobile smartphone applications, individual performance of mental health monitoring can be obtained in an in-the-moment and discrete manner. This could contribute to earlier intervention, crucial for mitigating the severity and long-term impact of depression. Preliminary predictions may also assist health practitioners in triaging patients at high risk and making optimal use of mental health services. Solving undiagnosed depression leads to overall productivity gains, lower costs for health care, and increased quality of life in social terms. Early detection could substantially enhance performance in school and at work among students and working adults. In the end, such mechanisms could help reduce suicide rates and reduce the fiscal burden that goes with mental illness for government and institutions. But there are also concerns about ethics and privacy to be weighed. Misinterpreted or misapplied predictions could further stigmatize or cause fear. To avoid harm, it is critical to ensure models are trained on diverse and unbiased data and are used together with human judgment.

5.2 Impact on Environment

Although the first goal of such a depression prediction system is healthcare related, it doesn't mean that the environmental impacts of a system cannot be considered. Like all ML applications, environmental impact is primarily driven by the computational resources consumed in model training and deployment. Training and development of machine learning models, in particular large or deep learning models, requires significant computational resources. This demand often leads to high-energy consumption, particularly when data centers are powered by nonrenewables. Training depression prediction models again and again, or at scale, has the potential to generate larger carbon emissions if poorly managed. Responsible ML processes, however, can mitigate this impact. Using light algorithms, cloud services certified with green energy and optimising code can reduce the impact on the environment drastically. Prediction models for depression (compared to other high-compute ML tasks – picture generation, large language models) are sometimes text/survey/structured data-based in an even less (computational) demanding manner. On the one hand, these systems may also cultivate environmentally sustainable behaviour in the sense that they favour psychological well-being. Studies have shown that people with higher mental health scores are more likely to engage in pro-environmental behaviors, such as recycling, curbing consumption and donating to environmental programs. Thus increased mental health promotes behavior that is environmentally friendly by virtue of positive spillover. In addition, those systems could reduce vehicle-related carbon-dioxide emissions—particularly in remote, rural pockets of the nation, where it’s often necessary to travel long distances for mental health care—if they cut the need for face-to-face sessions.

5.3 Ethical Aspects

Ethical considerations There are numerous ethics issues pertaining to employing machine learning (ML) for the prediction of depression that need to be considered for the ethical development and deployment. Data privacy and security: Mental health data is sensitive. However, in the presence of user data (e.g., survey responses, social media activities or medical data), privacy issues arise. Strong data encryption, anonymization, and secure storage procedures are needed to safeguard user identity and prevent abuse. Consent: Participants need to know how their information will be gathered, employed, and analyzed. They need to actively consent to participate. Hence we will safeguard the concern for individual freedom and transparency. Fairness and Bias: Data that is used to train a machine learning model can only be as unbiased as the model. The model may make biased or distorted predictions if the training data does not include a variety of ages or is subject to social biases, such as multiple generations of men and women or ages and ethnic backgrounds. An ethical deployment is facilitated by a range of fair-aware data and algorithms. False Positives, False Negatives: Bad predictions can have catastrophic consequences. In order to minimize this risk, the predictions do not replace final decisions by the clinician and the reader should always consult a professional before making an interpretation of the prognostic output. A false negative could cause someone to miss the treatment they need, while a false positive could cause undue anxiety or stigma. Labels and Stigmatization: Whether a label of ‘at risk’ for depression can have unintended social or stigmatizing effects is not known. Predictions should be sensitive as well as confidential to be dealt by systems. Responsibility: There should be some accountability of institutions and developers in how the system is used, and to ensure that it operates according to ethical and legal norms (in a general sense such as HIPAA or GDPR).

5.4 Sustainability Plan

In order to ensure the long-term effectiveness and relevance of the Depression Prediction with Machine Learning system, a clear sustainability plan is required. Such considerations may involve developing user engagement, mitigating environmental effects, ensuring ethical usage, and maintaining model accuracy. Model Upkeep: In order to keep up with changes in language use and patterns of mental health across time, the model must be updated frequently (for training on new dataset) and maintained (exist policy for the maintenance of the model). Adding actual use case and feedback loop data (with user consent) can also mitigate bias, improve prediction accuracy. Ethical/Legal: Compliance Sustainability is being compliant with data protection laws such as the U.S. government's HIPAA or the European Union's GDPR. There is a need for periodic monitoring for ethical compliance and clear informed consent of the users at present and in the near future. Responsible-AI methodologies will preserve trust, prevent misuse. Environmental considerations: Our system should use cloud platforms powered by renewable energy, and energy efficient ML algorithms to reduce its carbon footprint. The model training should run using world's smallest computer and should be optimized in resource consumption at up performance level. Community and Stakeholder Engagement: In order to maximize the value of the tool and maintain clinical relevance, the tool will be developed with input from patients, researchers, and mental health professionals. Community feedback could inform betterment, and collaborations with health care providers or NGOs could encourage greater usage. Access and Inclusivity: The system should be accessible in various languages and through a wide range of devices for sustainable impact. Interfaces that encourage digital divides need to be easy to use and flexible for a range of people.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

Given the availability of a Kaggle dataset featuring diverse behavioral, psychological, and demographic features, this study aimed to explore the utility of machine learning algorithms in predicting depression. The aims of this study were to determine the best machine learning model, feature selection method, and to explore the dataset, for predicting depression. Recursive Feature Elimination (RFE) was employed as a critical part of the feature selection process to reduce the dimensions of the dataset and to make the model computationally more efficient. The study results show that ensemble classifiers (XGBoost, Gradient Boosting, and AdaBoost) achieved higher AMCC compared with conventional classifiers (Decision Tree, K-Nearest Neighbors (KNN), and Logistic Regression). These models have good prediction performance (for example, the highest accuracy of classification was 95%). Results on models with and without RFE suggested that removal of irrelevant dimensions had no impact on accuracy, which indicated that feature selection could be the effective way to simplify the dataset with no loss of performance. That there existed a class imbalance in the sample such that, most participants were labelled as Not Depressed (0) was also addressed in the study. Features importance examination showed that important factors to predict depression were work pressure, financial hardship, job satisfaction, sleep hours, and family history of mental health. In general, the results demonstrate that machine learning presents a feasible and scalable method for the assessment of mental health. The integration of feature selection, class balancing and the use of ensemble models ensures the accuracy and reliability of the recognition of depression. Based on these findings, machine learning could be integrated into AI-supported diagnostic models, and early intervention systems, and mental health screening tools, making a significant contribution to the promotion of mental health care and health knowledge.

6.2 Limitations

There are, however, some limitations that should be emphasized, despite the encouraging results prospectively. One of the drawbacks is data imbalance, as most of the population in the dataset was labelled as Not Depressed (0)), which gave the model a bias when training on the model. A limitation is the loss of data for key factors like Academic Pressure, Study Satisfaction, Job Satisfaction. To address this issue, missing values were deleted or imputed by performing descriptive statistical techniques such as mean and median imputation. While this may not preserve the true trend for the underlying data as well, more complex methods such as deep learning-based imputation techniques or multiple imputations will help ensure that good data quality is maintained. The study also drew on data from structured surveys; however these might not be enough to capture the complexities of mental health. There are psychological, environmental, and genetic factors and much of it cannot be picked up by self-reported survey. In the future, these rates will perhaps be supplemented with real-time behavioral data, wearable device readings and even social media mood analysis, for a more complete picture of mental health. In addition, the RFE was successful in reducing the dimension of the dataset without a significant loss in accuracy, although some of the features selected may not be general to every demographic. Feature selection can be effective depending on the dataset used and the results may be different in different cultural, socioeconomic or geographic contexts. Lastly, this study focused on machine learning classification models, and did not consider deep learning approaches like neural networks or transformer models, which could have boosted predictive performance. But deep learning requires a lot of processing power and vast amounts of data, neither of which were in the scope of this study.

6.3 Future Work

Based on the limitation and results of the study, some suggestions for future research to improve the reliability, generalizability, and practicality of the machine learning methods in the prediction of depression can be recommended. An immediate goal for improvement lies in larger and more diverse data set inclusion. So sosomething worth noting here is the dataset the researchers used in the study, which is from Kaggle, may not accurately represent various socioeconomic classes, demographic groups, or cultural backgrounds. Longitudinal mental health records, multiregional survey and true clinical data are needed for generative models in future research. One final important suggestion is move to exploring state-of-the-art deep learning techniques, such as deep neural networks, transformers, and mixed hybrid deep learning models that can potentially capture more complex nonlinear inter-dependencies in mental health data. Multimodal data fusion should also considered in the future work. And by in addition using behavioral signals from smartphone use patterns, sleep tracking, physical activity, or even EEG measurements (electroencephalograms) researchers will be able to sharpen their predictions even more. Combine these disparate data sources with machine learning models and you're starting to be able to get a fuller picture of what someone's mental health might look like. Secondly, further work is required for promoting dimensionality reduction or selecting features. Although Recursive Feature Elimination (RFE) was shown to be effective in this study, more sophisticated feature selection methods, such SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Modelagnostic Explanations), may be able to provide a better understanding of the most important predictors of depression. Such methods would aid create interpretable AI models to keep projections transparent and comprehensible to mental health experts Finally, future research should consider the ethical and privacy implications to predict mental health using AI. For diagnostic decisions, healthcare applications, and mental health chatbots, the role of machine learning in ensuring data privacy, fairness and accountable AI practices becomes imperative. To ensure the privacy of mental health sensitive information, researchers should investigate privacy-preserving machine learning techniques (e.g., federated learning, differential privacy).

REFERENCES

1. A. Rahman, A. A. Ananna, S. Fahad and M. S. Mia, "Prediction of Depression and Anxiety on University Students in Bangladesh Using Machine Learning," *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, Dhaka, Bangladesh, 2023, pp. 1-6, doi: 10.1109/STI59863.2023.10465126.
2. M. H. Amirhosseini, A. L. Ayodele and A. Karami, "Prediction of Depression Severity and Personalised Risk Factors Using Machine Learning on Multimodal Data," *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, Varna, Bulgaria, 2024, pp. 1-7, doi: 10.1109/IS61756.2024.10705185.
3. N. K. Trivedi, R. G. Tiwari, D. Witarsyah, V. Gautam, A. Misra and R. A. Nugraha, "Machine Learning Based Evaluations of Stress, Depression, and Anxiety," *2022 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, Bandung, Indonesia, 2022, pp. 1-5, doi: 10.1109/ICADEIS56544.2022.10037336.
4. G. H. Suhas, L. Suraj, J. Varun, D. V. Veda and H. S. Jayanna, "Machine Learning Approaches for Detecting Early-Stage Depression using Text," *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, Mysuru, India, 2021, pp. 106-110, doi: 10.1109/ICEECCOT52851.2021.9707950.
5. *B. Mahesh and M. Amanullah, "Prediction of Depression using the KNeighbors-Classifier Algorithm Compared for Improved Accuracy with Support Vector Machine," 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), Chennai, India, 2023, pp. 1-4, doi: 10.1109/ICCEBS58601.2023.10449307.*
6. M. S. Keya and A. Han, "A Performance Analysis of Depression Ratio using Machine Learning Approaches," *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2022, pp. 215-219, doi: 10.1109/ICAIS53314.2022.9742757.
7. K. Shah, U. Patel and Y. Kumar, "Machine Learning-Based Approaches for Early Prediction of Depression," *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, Bangalore, India, 2024, pp. 1-7, doi: 10.1109/IITCEE59897.2024.10467234.
8. P. Nison, P. Vuttipittayamongkol, P. Boonyapuk and K. Kemavuthanon, "A Machine Learning Approach for Depression Screening in College Students Based on Non-Clinical Information," *2023 International Conference On Cyber Management And Engineering (CyMaEn)*, Bangkok, Thailand, 2023, pp. 413-417, doi: 10.1109/CyMaEn57228.2023.10051001.
9. P. P. Shinde, V. P. Desai and K. S. Oza, "A Data Driven Mental Health Analysis Using Machine Learning Techniques," *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Coimbatore, India, 2024, pp. 1665-1670, doi: 10.1109/ICoICI62503.2024.10696584.
10. H. Dhawale, D. Thakare, N. C. Morris, R. Agrawal and C. Dhule, "The Prediction of Mental Health Using Machine Learning," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10726276.

11. N. Bagga, P. Vashistha and P. Yadav, "Predicting Depression from Social Networking Data using Machine Learning Techniques," *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India, 2021, pp. 849-854, doi: 10.1109/ICAC3N53548.2021.9725421.
12. H. Abdulla, M. Maalouf and H. F. Jelinek, "Machine Learning for the Prediction of Depression Progression from Inflammation Markers," *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Sydney, Australia, 2023, pp. 1-4, doi: 10.1109/EMBC40787.2023.10340436.
13. J. M. Shanthi, H. Sharma, P. S. Rao, G. A. S. Thomas and A. Gupta, "Machine Learning Technique Enabled Learning Methodology for Human Depression Prediction," *2024 Second International Conference on Advances in Information Technology (ICAIT)*, Chikkamagaluru, Karnataka, India, 2024, pp. 1-6, doi: 10.1109/ICAIT61638.2024.10690553.
14. S. Rosaline, R. S. Kaavya Varshitha, K. NV and K. Spoorthi, "A Novel Approach to Early Depression Prediction and Estimation with EEG Signals," *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, 2022, pp. 1240-1243, doi: 10.1109/ICAISS55157.2022.10010875.
15. S. Singh, H. Gupta, P. Singh and A. P. Agrawal, "Comparative Analysis of Machine Learning Models to Predict Depression, Anxiety and Stress," *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India, 2022, pp. 1199-1203, doi: 10.1109/SMART55829.2022.10047752.
16. U. R. K and K. Latha, "Early Depression Prediction and Estimation with EEG Signals using Machine Learning Algorithm," *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, Chennai, India, 2022, pp. 1-5, doi: 10.1109/IC3IOT53935.2022.9767986.
17. L. K. Xin and N. b. A. Rashid, "Prediction of Depression among Women Using Random Oversampling and Random Forest," *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, Taif, Saudi Arabia, 2021, pp. 1-5, doi: 10.1109/WiDSTaif52235.2021.9430215.

Early Depression Detection Using Machine Learning .pdf

ORIGINALITY REPORT

9%	7%	4%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Progressive Computational Intelligence, Information Technology and Networking", CRC Press, 2025 Publication	1%
3	"Machine Learning for Cyber Physical System: Advances and Challenges", Springer Science and Business Media LLC, 2024 Publication	<1%
4	Lalit Prasad, Anatoliy Goncharuk, Teddy Fauzi, Hanna Doroshuk, Sri Sundari. "Sustainable Smart Technology Businesses in Global Economies - Proceedings of International Conference on Sustainable Smartech Businesses and SMEs across Global Economies", Routledge, 2025 Publication	<1%
5	jcsdf.nfsu.ac.in Internet Source	<1%
6	R. N. V. Jagan Mohan, Vasamsetty Chandra Sekhar, V. M. N. S. S. V. K. R. Gupta. "Algorithms in Advanced Artificial Intelligence", CRC Press, 2024 Publication	<1%
7	www.letstalkcounselingandservices.com Internet Source	<1%