STRESS AND ANXIETY EARLY DETECTION WITH MACHINE LEARNING APPROACH USING SURVEY DATA

BY

BANDHAN GHOSH

ID: 201-15-14196

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Abu Kaisar Mohammad Masum
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled **"Stress and anxiety early detection with machine learning approach using survey data"**, submitted by **Bandhan Ghosh** 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 **13-07-2024**.

BOARD OF EXAMINERS

Dr. Md. Ismail Jabiullah (MIJ) — **Board Chairman**
Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Abdus Sattar (AS) — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Ms. Zahura Zaman (ZZ) — **Internal Examiner 2**
Lecturer
Department of CSE
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 **Abu Kaisar Mohammad Masum, Lecturer, 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:

For

Abu Kaisar Mohammad Masum
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Mohammad Jahangir Alam
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

Bandhan Ghosh

Bandhan Ghosh
ID: 201-15-14196
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty God for His divine blessing in making it possible to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Mr. Abu Kaisar Mohammad Masum**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our 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 stage have made it possible to complete this project.

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

We would like to thank our entire course mate 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 my parents.

ABSTRACT

Stress and anxiety can have profound impacts on an individual's mental and physical well-being, making their early detection crucial for timely intervention and support. This study investigates the application of machine learning for detecting stress and anxiety levels from survey data, which is crucial for timely intervention and support. A comprehensive experimental analysis evaluated various algorithms, including Support Vector Machines (SVM), XGBoost, CatBoost, ensemble methods like Random Forest, and traditional techniques like Logistic Regression. The experimental setup utilized powerful computational resources and employed tools like NumPy, Pandas, and Scikit-learn. Extensive data preprocessing and feature engineering ensured data quality. Results revealed SVM, XGBoost, and CatBoost as top performers, achieving highest accuracy and F1 scores, demonstrating effectiveness in capturing complex stress and anxiety patterns. However, challenges like data quality, feature selection, individual variability, and interpretability were encountered. The study highlights machine learning's potential for stress and anxiety detection, providing insights into algorithm selection and performance trade-offs. Future work may focus on incorporating domain knowledge, addressing ethical considerations, and deploying the model for improved mental health support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

In the present era, stress and anxiety have become greatly prevalent in our everyday lives due to the fast-paced nature of modern society. These mental health issues can have severe negative impacts on an individual's overall well-being and productivity. Early detection and management of stress and anxiety is thus of utmost importance. Traditional methods of diagnosis rely heavily on subjective evaluations by mental health professionals, which can often be time-consuming and prone to bias.

This research is to develop an objective, data-driven approach to detect stress and anxiety levels in individuals using machine learning techniques on survey response data. A comprehensive set of 13 questions related to common symptoms of stress and anxiety such as lack of focus, irritability, risk-taking behavior etc. was formulated. Responses to this survey from a large number of participants were collected and used to train and evaluate various machine learning classification models.

A total of 13 different machine learning algorithms were applied, including Support Vector Machines (SVM), XGBoost, CatBoost, Gradient Boosting, Random Forest, Extra Trees, Logistic Regression, Decision Tree, Bagging Classifier, Naive Bayes, Ridge Classifier, AdaBoost, Stochastic Gradient Descent and Passive Aggressive Classifier. The core objective was to identify the model(s) that could most accurately predict whether a person was suffering from high levels of stress and anxiety based solely on their survey responses.

The study holds promising potential to contribute towards an automated screening system that can efficiently identify at-risk individuals. This could allow for timely intervention and treatment, consequently improving overall mental health outcomes in the population. The following sections elaborate on the methodology, results, analysis and conclusions drawn from this novel application of machine learning for stress and anxiety detection.

1.2 Motivation

Stress and anxiety are highly prevalent mental health issues that affect a significant portion of the global population. In Bangladesh specifically, studies have shown an alarming rise in the number of people suffering from these conditions in recent years. The fast-paced, competitive nature of modern urban life, coupled with socio-economic pressures, have contributed to this increasing trend.

Untreated chronic stress and anxiety can have severe detrimental impacts on an individual's physical and mental well-being. They are linked to numerous adverse health consequences such as cardiovascular diseases, diabetes, depression, and even increased risk of mortality. Additionally, these conditions can substantially impair productivity, interpersonal relationships, and overall quality of life.

Early detection and management of stress and anxiety are thus crucial from both a public health and economic standpoint. However, traditional diagnostic methods involving consultations with mental health professionals can often be time-consuming, expensive, and suffer from subjective biases. There is a critical need for objective, data-driven approaches that can help screen large populations efficiently and identify at-risk individuals.

The motivation behind this research stems from the immense potential of machine learning to provide such a solution. By training models on survey data capturing common symptoms and behaviors associated with stress and anxiety, we aim to develop an automated system capable of accurately detecting these conditions. Such a tool could revolutionize mental healthcare accessibility and delivery, especially in resource-constrained settings like Bangladesh.

1.3 Rationale of the Study

The rationale behind conducting this study is multifaceted, driven by the pressing need to address the growing prevalence of stress and anxiety disorders, as well as the limitations of current diagnostic approaches. The key rationales are as follows:

Early Detection and Intervention: Stress and anxiety, if left unaddressed, can escalate into more severe mental health issues and contribute to a range of adverse physiological outcomes. By enabling early detection through an automated screening system, this study aims to facilitate timely intervention and treatment, thereby preventing further deterioration of an individual's condition.

Objective and Scalable Screening: Traditional diagnostic methods heavily rely on subjective evaluations by mental health professionals, which can be time-consuming, expensive, and susceptible to biases. The machine learning approach proposed in this study offers an objective and scalable screening solution that can potentially reach a larger population, especially in resource-constrained settings.

Data-Driven Decision Support: The study aims to leverage the power of data and machine learning algorithms to provide an evidence-based decision support system for mental health professionals. The trained models can aid in identifying patterns and risk factors associated with stress and anxiety, potentially enhancing diagnostic accuracy and treatment planning.

Accessibility and Cost-Effectiveness: Mental healthcare services are often inaccessible or unaffordable for a significant portion of the population, particularly in developing countries like Bangladesh. The proposed automated screening system has the potential to increase accessibility and reduce the overall cost of mental healthcare delivery, thereby promoting greater inclusivity.

Advancement of Machine Learning Applications: This study contributes to the advancement of machine learning applications in the domain of mental health, paving the way for future research and development in this field. The insights gained from this research can inform and inspire similar endeavors, ultimately leading to improved mental healthcare practices.

By addressing these rationales, the study has the potential to make a significant impact on mental health outcomes, healthcare accessibility, and the overall well-being of individuals and communities.

1.4 Research Questions

The primary research questions that this study aims to address are:

- Can machine learning models effectively detect stress and anxiety levels in individuals based solely on their responses to a survey questionnaire capturing common symptoms and behaviors associated with these conditions?
- Which specific machine learning algorithm or ensemble of algorithms performs best in accurately classifying individuals as suffering from high or low levels of stress and anxiety?
- What are the key features or survey questions that hold the most predictive power in determining an individual's stress and anxiety levels according to the trained models?

- How does the performance of the developed models compare to traditional diagnostic methods and human evaluations in terms of accuracy, efficiency, and scalability?
- Can the trained models identify any specific demographic or socio-economic factors that may contribute to an increased risk of developing stress and anxiety disorders?
- What are the potential real-world applications and implications of deploying such an automated stress and anxiety detection system, particularly in resource-constrained settings like Bangladesh?
- How can the insights gained from this study inform and advance the field of machine learning applications in mental healthcare, paving the way for future research and development?

By answering these research questions, the study aims to provide a comprehensive understanding of the feasibility, performance, and potential impact of leveraging machine learning techniques for the early detection and screening of stress and anxiety disorders.

1.5 Expected Output

The expected output of this research study can be categorized into two main aspects: the development of accurate machine learning models for stress and anxiety detection, and the insights gained from the analysis of these models.

Model Development:

- Trained and optimized machine learning models capable of classifying individuals as suffering from high or low levels of stress and anxiety based on their survey responses.
- Identification of the best-performing algorithm(s) or ensemble techniques that achieve the highest accuracy in stress and anxiety detection.
- Evaluation metrics such as precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) for each model, allowing for a comprehensive assessment of their performance.

Insights and Analysis:

- Interpretation of the trained models to understand the key features or survey questions that hold the most predictive power in determining an individual's stress and anxiety levels.
- Identification of potential demographic, socio-economic, or lifestyle factors that may contribute to an increased risk of developing stress and anxiety disorders, based on the patterns learned by the models.
- Comparative analysis of the developed models' performance against traditional diagnostic methods and human evaluations, in terms of accuracy, efficiency, and scalability.
- Assessment of the potential real-world applications and implications of deploying an automated stress and anxiety detection system, particularly in resource-constrained settings like Bangladesh.
- Recommendations and guidelines for the effective implementation and integration of the developed models into existing mental healthcare practices and systems.

Additionally, the study is expected to contribute to the broader field of machine learning applications in mental healthcare by providing a comprehensive framework and methodology for leveraging survey data and predictive modeling techniques in the early detection and screening of mental health conditions.

1.6 Project Management and Finance

This research was undertaken as an academic endeavor without external financial support. Costs were kept to a minimum by leveraging freely available tools and resources.

Data Collection: Data collection was facilitated through various channels including Google Forms, offline surveys, and questionnaires, ensuring a diverse and comprehensive dataset.

Computing Infrastructure: Computational tasks were performed using VS Code on local machines, utilizing available computing resources efficiently.

Software Tools: Open-source libraries such as NumPy, Pandas, scikit-learn, and Seaborn were extensively utilized for data analysis and model development, enabling cost-effective and robust implementation.

Team Members: The core research was conducted solely by the principal investigator, eliminating the need for additional team members or associated salaries.

The minimal expenses incurred were managed through personal funding without external sponsorship. By maximizing the utilization of free online resources and tools, the research progressed effectively within the constraints of a limited budget.

1.7 Report Layout

This report is structured into six chapters as outlined below:

Chapter 1: Introduction

This chapter provides an overview of the research by introducing the topic, motivations, objectives, and expected outcomes. It outlines the research questions, project management, and the structure of the report.

Chapter 2: Literature Review

This chapter reviews relevant literature and terminology, analyzing existing works related to the problem domain and identifying key gaps. It defines the scope of the study and discusses any challenges encountered.

Chapter 3: Methodology

Chapter 3 contains details of the research methodology, describing the datasets used, statistical analysis techniques applied, proposed models, experimental setup, and implementation requirements.

Chapter 4: Experimental Results and Analysis

This chapter presents detailed experimental results, analysis, and discussion for each phase of the study. It evaluates and interprets the outcomes of different models used in the research.

Chapter 5: Societal Impacts and Considerations

Chapter 5 analyzes the cultural, societal, ethical, and sustainability impacts of the research, considering possible applications and societal considerations.

Chapter 6: Conclusion and Future Work

The final chapter summarizes the overall research findings, draws conclusions, highlights limitations, and provides recommendations for future work and research directions.

CHAPTER 2
Background

2.1 Terminologies

Confusion Matrix

A table that summarizes the performance of a classification model by representing the correct and incorrect predictions made by the model.

True Positive (TP)

The number of instances that were correctly classified as positive by the model.

True Negative (TN)

The number of instances that were correctly classified as negative by the model.

False Positive (FP)

The number of instances that were incorrectly classified as positive by the model.

False Negative (FN)

The number of instances that were incorrectly classified as negative by the model.

Log Loss

A metric that quantifies the accuracy of a classifier by penalizing false classifications. It works well for multi-class classification problems and provides more nuance than accuracy.

F1-Score

A measure that combines precision and recall into a single metric, providing a balanced evaluation of a model's performance. It is calculated as the harmonic mean of precision and recall.

$$F1Score = \frac{2 \times Precission \times Recall}{Precission + Recall}$$

Precision

The proportion of correctly predicted positive instances out of all instances predicted as positive. Precision measures the accuracy of positive predictions made by the model. It is particularly useful when the cost of false positives is high.

$$Precision \ = \frac{TP}{TP + FP}$$

Recall

The proportion of correctly predicted positive instances out of all actual positive instances. Recall measures the model's ability to find all positive instances. It is important when the cost of false negatives is high.

$$Recall \ = \frac{TP}{TP + FN}$$

Accuracy

The fraction of instances that were correctly classified by the model out of the total number of instances.

$$Accurecy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FlaseNegative}$$

2.2 Related Works

The literature on stress and anxiety underscores their multifaceted nature, particularly in nursing research focusing on psychobiologic variables, life events' influence, and the stress response's interactional model. Stress, difficult to define, manifests as anxiety, signaling the initiation of the stress response, which triggers immediate psychophysiologic reactions lasting up to three weeks post-stressor confrontation, potentially impacting immunocompetence via neuroendocrine alterations. Interventions encompass pharmacotherapy, psychotherapy, and behavioral techniques, reflecting the proliferation of nursing research on stress since the 1980s. Coping strategies spanning behavioral, physical, cognitive, and emotional realms emerge as crucial in nursing interventions for stress management.[1]

Understanding the intricate relationship between stress and anxiety is crucial for managing psychopathologies like PTSD. This review highlights shared neural mechanisms in pre-clinical models, focusing on basolateral amygdala circuits, locus coeruleus inputs, mitochondrial function in the nucleus accumbens, and hypothalamic corticotropin-releasing hormone neurons. These findings underscore the overlap in processing stress and anxiety, offering insights into novel therapeutic approaches.[2]

This study explores how a social stressor impacts anxiety in humans by examining startle response facilitation in the dark. Results indicate that stress significantly increases anxiety, as evidenced by heightened startle response and subjective distress. Physiological measures also show elevated stress responses. These findings support the notion that stress exacerbates anxiety, with potential implications for anxiety disorder management.[3]

Recent neuroimaging research focuses on unraveling the brain circuits behind anxiety disorders, drawing from studies of fear circuits in animals and brain responses to emotional stimuli in healthy individuals. This review synthesizes findings, indicating heightened amygdala activation in response to disorder-relevant stimuli across post-traumatic stress disorder, social phobia, and specific phobia. Additionally, the insular cortex shows increased activation in many anxiety disorders, while post-traumatic stress disorder exhibits reduced responsiveness in the rostral anterior cingulate cortex and adjacent ventral medial prefrontal cortex. Future research aims to clarify the role of fear circuit components, understand functional abnormalities, integrate with neurochemistry studies, and utilize neuroimaging to predict and assess treatment response.[4]

Isatin, an endogenous indole present in mammalian brain and peripheral tissues, exhibits diverse behavioral and metabolic effects, with its concentration in blood exceeding 1 μM and varying in tissues from < 0.1 to 10 μM. Elevated by stress, it demonstrates anxiogenic properties at lower doses and sedative effects at higher doses. Isatin's main in vitro actions include antagonizing atrial natriuretic peptide (ANP) function and nitric oxide (NO) signaling. This review discusses isatin's role in animal models and limited human studies, along with molecular mechanisms of action. Potential pharmacological applications are explored, suggesting isatin antagonists as anxiolytic agents and isatin agonists for HPA axis activation.[5]

Two experiments built on prior research by MacLeod and Mathews (1989) to explore whether cognitive bias for threat information is affected by state or trait anxiety. Using stress manipulation procedures, the studies examined high and low trait anxious individuals in color-naming and attention tasks. Results revealed that high stress led to selective processing of threat stimuli regardless of trait anxiety levels. There was no consistent evidence of a cognitive bias associated with trait anxiety, and stress effects were not mediated by state anxiety. This suggests trait factors may not modify attentional biases during acute stress but could under prolonged stress.[6]

This article explores the emerging focus on academic stress among UK schoolchildren, addressing conceptual and methodological challenges. It highlights ambiguity in terminology usage, particularly 'stress', 'anxiety', and 'worry', and the lack of clear definitions for 'examination stress' and 'academic stress'. Additionally, it critiques the bias towards quantifying and measuring stress and anxiety, suggesting the need for alternative approaches.[7]

Dentists experience considerable professional stress, starting from dental school, impacting their personal and professional lives. This stress can lead to burnout, anxiety disorders, and depression, exacerbated by clinical practice demands and common personality traits among dentists. However, treatment and prevention strategies exist to help manage these challenges. Dentists must prioritize their physical and mental well-being for fulfilling personal and professional lives, understanding the implications of stress for effective practice management.[8]

The brain's high oxygen consumption, modest antioxidant defenses, and lipid-rich composition make it susceptible to redox imbalances, leading to oxidative damage and nervous system impairment. Recent research has implicated oxidative stress in conditions like depression, anxiety disorders, and high anxiety levels. These findings have spurred further investigation into the link between oxidative status and normal anxiety, as well as the potential role of oxidative stress in anxiety genesis. This review explores recent discoveries in this area, discussing differing perspectives and methodological approaches aimed at elucidating the causal relationship between oxidative and emotional stress.[9]

Work-related stressors, called stress-producing environmental circumstances (SPECs), can lead to anxiety, termed work-related strain. Social support is often suggested to alleviate this strain by directly reducing anxiety, lessening SPECs' impact on anxiety, or weakening SPECs themselves. Future research could investigate these relationships further by designing studies for stronger causal inference, testing theoretical hypotheses, and exploring cross-cultural differences in social support's effects on anxiety in the context of work-related SPECs.[10]

[Authors] explored lasting behavioral effects of acute stressors in male rats, using an approach/avoidance paradigm. Social defeat or electric shocks led to significant reductions in exploration of a compartment containing an unfamiliar rat, lasting over 5 days for social defeat and 5-10 days for electric shocks. Stress inhibition was more pronounced in the presence of a stimulus rat. The test's repeatability was shown, and chlordiazepoxide pretreatment alleviated stress-induced deficits, indicating potential for assessing anxiolytic compounds. This research informs understanding of stress-induced anxiety and potential pharmacological interventions.[11]

The impact of psychological stress on immune function was investigated in 38 medical students facing academic examination stress by [Authors]. Their study revealed that stress significantly increased production of pro-inflammatory cytokines TNF-α, IL-6, IL-1Ra, IFN-γ, and immunoregulatory cytokines IL-10. Those perceiving high stress showed elevated levels of TNF-α, IL-6, IL-1Ra, and IFN-γ, while high anxiety was associated with increased IFN-γ and decreased IL-10 and IL-4, suggesting a T-helper-1-like response. This highlights the influence of stress perception and anxiety on immune reactivity in humans.[12]

In a study of 38 medical students facing academic stress, [Authors] found that psychological stress significantly increased production of pro-inflammatory cytokines (TNF-α, IL-6, IFN-γ) and immunoregulatory cytokines (IL-10). Students perceiving higher stress levels showed elevated TNF-α, IL-6, IL-1Ra, and IFN-γ, while those with increased anxiety had higher IFN-γ and lower IL-10 and IL-4 levels, suggesting a T-helper-1-like response. These findings highlight the impact of stress perception and anxiety on immune function, underscoring the role of psychological factors in shaping immune responses to stress in humans.[13]

Considerable preclinical evidence supports the association between noradrenergic brain systems and stress and anxiety-related behaviors. The locus coeruleus, housing the majority of noradrenergic neurons, exhibits increased firing and norepinephrine release in response to stressors, influencing various brain regions. Chronic stress leads to long-term alterations in locus coeruleus activity and norepinephrine release, impacting fear-related behaviors and neural mechanisms like sensitization and fear conditioning. These findings are pertinent to understanding psychiatric disorders like panic disorder and PTSD, where noradrenergic dysfunction is implicated.[14]

Anxiety disorders, depression, and alcohol use disorder often co-occur and are linked to oxidative stress, resulting from an imbalance in reactive oxygen species and defense mechanisms. This stress damages macromolecules, affecting crucial cellular functions, particularly in the brain. Recent research has examined oxidative stress markers in patients and animal models of these disorders. Despite interest, detailed mechanisms connecting oxidative stress to disease remain unclear, warranting further study for potential pharmacological interventions.[15]

Examination stress is a significant concern for college students, impacting their academic performance. This study investigated stress levels among Arts, Science, and Commerce students in undergraduate and postgraduate programs. Results showed a correlation between examination stress and anxiety, with Arts students experiencing the highest levels, followed by Commerce students. However, no significant difference was found between undergraduate and postgraduate students. These findings highlight the need for targeted interventions to alleviate examination stress and promote student well-being.[16]

This study explores stress, anxiety, and periodontal health in 79 patients categorized by probing pocket depth (PPD). Stress and anxiety were assessed using the SSI, SRRS, and STAI, while clinical measures like PI, GI, PPD, and CAL were collected alongside medical and socioeconomic data. While no significant differences were found in stress or anxiety levels between groups, individuals with higher trait anxiety showed a stronger association with moderate CAL and PPD, suggesting a potential link to periodontal disease susceptibility.[17]

Anxiety disorders, affecting 18% of the US population, may involve oxidative stress-induced damage to proteins, lipids, and nucleic acids. Dysfunction in cellular proteolytic systems, such as the proteasome and Lon protease, may lead to toxic protein accumulation in the brain, potentially influencing the severity of mood disorder symptoms. Understanding these mechanisms could inform the development of new treatments.[18]

A study of 1,415 men, all 26 years old and part of the National Survey of Health and Development, explored nervous strain at work, predisposition to anxiety, and job activities. It found that high-level job roles were associated with more reported strain compared to manual work. Both anxiety predisposition and specific job factors (such as supervising, teaching, contact with people, driving, and skilled machine work) equally contributed to reported strain, regardless of job level. Surprisingly, there was little evidence that anxious individuals held more stressful jobs. The study concluded that predisposition to anxiety and specific job stressors independently influenced reported nervous strain at work.[19]

This study of 71 Nursing students aimed to describe their social-demographic data and correlate stress and anxiety levels using Vasconcelo's List of Stress Symptoms and Spielberger's Inventory of Trait and State Anxiety. Most students were female (90.44%) with an average age of 28.87 years. Stress levels were predominantly high, with 64.79% reporting high scores, while 43.66% reported high scores in Trait and State Anxiety. Positive correlations were found between Life Stress Scale (LSS) scores and Trait-Anxiety ($r=0.656$, $p=0.000$), as well as between LSS scores and State-Anxiety ($r=0.512$, $p=0.000$).

These findings highlight significant stress and anxiety levels among Nursing students, underscoring the need for further research.[20]

This fMRI study investigated gender differences in stress-induced anxiety responses in 96 healthy individuals with similar trait anxiety levels. Men showed heightened activity in the caudate, cingulate gyrus, midbrain, thalamus, and cerebellum, while women exhibited increased activity in the posterior insula, temporal gyrus, and occipital lobe. Regression analyses revealed gender-specific neural patterns, with women displaying positive associations between medial prefrontal–parietal cortex activity and anxiety, and men showing negative associations. These findings shed light on gender-specific neural mechanisms underlying stress-induced anxiety, potentially informing tailored treatment strategies for stress-related clinical disorders.[21]

Cannabinoid agonists have complex effects on anxiety, influenced by dose, context, and interactions with CB1 and non-CB1 receptors, as well as other systems like CRH, GABAA, cholecystokinin, opioids, and serotonin. Brain regions with high CB1 receptor densities, such as the amygdala, hippocampus, and cortex, play a key role in emotional regulation. Mice lacking CB1 receptors exhibit anxiety-like and depressive-like behaviors, while blocking CB1 receptors induces anxiety in rats. Inhibiting anandamide metabolism produces anxiolytic effects. These findings underscore the endocannabinoid system's importance in emotional regulation and its potential as a target for anti-anxiety therapy.[22]

The panic response, crucial for coping with threats, involves behavioral, cardiorespiratory, and endocrine changes. Identified in the posterior hypothalamus, particularly in the perifornical (PeF) and dorsomedial (DMH) regions, this area regulates circadian rhythms. Orexin (ORX), discovered in 1998, primarily regulates wakefulness and energy balance and is linked to narcolepsy. Recent studies also connect ORX to depression and anxiety, suggesting its role in coordinating panic responses and its potential hyperactivity in pathological anxiety states.[23]

This study develops a framework using video-recorded facial cues to detect stress/anxiety states. An experimental protocol induced varied affective states (neutral, relaxed, and stressed/anxious) through different stressors. Analysis focused on eye-related events, mouth activity, head motion, and camera-based heart rate estimation. Features were selected and used for classification, achieving good accuracy in discriminating stress/anxiety from neutral states. Results suggest specific facial cues as reliable indicators of stress and anxiety.[24]

It uses information from the DASS-21 questionnaire to investigate machine learning algorithms for the prediction of stress, anxiety, and depression. Five algorithms were used to examine data from various populations; Random Forest produced the best results in terms of accuracy. The study emphasizes the potential of ML in improving mental health diagnosis and treatment efficiency by addressing class imbalances using the F1 score. To validate these results in larger populations, more research is required.[25]

In order to predict anxiety, depression, and stress using the DASS42 and DASS21 datasets, this study employs eight machine learning methods that are categorized into probabilistic, closest neighbor, neural network, and tree-based categories. The radial basis function network demonstrated the highest accuracy among hybrid classification algorithms, which outperformed standalone algorithms in terms of accuracy.[26]

Using a Likert scale questionnaire, 127 Indian engineering students' anxiety levels and effects were evaluated. Data validity (Pearson's correlation = 0.823) and reliability (Cronbach's alpha = 0.723) were verified by statistical testing. Anxiety levels were categorized by machine learning algorithms with accuracy rates of 71.05% for Decision Tree and Naive Bayes, 78.9% for Random Forest, and 75.5% for SVM.[27]

DASentimental is a neural network-powered tool that predicts depression (R = 0.7), anxiety (R = 0.44), and stress (R = 0.52) based on free associations using semantic memory and network distances. When it was applied to a clinical dataset containing 142 suicide notes, it matched the circumplex model of affect by properly predicting degrees of anxiety and depression. Future studies will examine how AI can identify emotional distress in texts.[28]

In the WESAD dataset, wearable sensors (ACC, ECG, BVP, TEMP, RESP, EMG, and EDA) provide multimodal data for the purpose of this paper's exploration of stress detection utilizing machine learning and deep learning. Machine learning techniques such as K-Nearest Neighbour, Random Forest, and AdaBoost obtained up to 81.65% and 93.20% accuracy, respectively, for three-class (amusement, baseline, stress) and binary (stress vs. non-stress) classifications. For the same classes, 84.32% and 95.21% accuracy were attained by deep learning techniques.[29]

This work introduces a machine learning pipeline that utilizes acoustic features from speech samples to predict the severity of depression, anxiety, and stress using a longitudinal dataset labeled with DASS-21 scores. According to preliminary findings, study participants in good health adhere to protocol more closely than those with mental health symptoms.

When trained using VGG-19 features, a one-dimensional convolutional neural network showed excellent accuracy in predicting the degree of stress, anxiety, and depression.[30]

In order to diagnose cognitive biases in subclinical depression and anxiety, this study created a diagnostic technique. To differentiate anxiety from depression, data from 125 participants was evaluated using a behavioral test battery and sophisticated machine-learning algorithms. When predicting symptomatic groups against control groups, the model had a 71.44% sensitivity and a 70.78% specificity. Important behavioral indicators that influenced these forecasts were found, which aided in the creation of more precise diagnostic instruments and tailored therapies.[31]

Using Internet of Things (IoT) devices to gather and evaluate real-time vital sign data, this eHealth system assists educators and students in managing stress. The technology, which is backed by fog computing, makes use of machine learning to forecast anomalies and provides immediate insights. When assessed using heart rate data, it demonstrates the possibility of individualized feedback and in-the-moment interventions in learning environments.[32]

Using a survey dataset that includes demographic data and the results of the DASS-42 psychometric test, we compare machine learning and ensemble learning techniques. Five severity levels are used by Random Forest, Decision Tree, SVM, AdaBoost, CatBoost, and XGBoost to categorize depression, anxiety, and stress. SVM fared better than the others, obtaining F1-measures for stress, anxiety, and depression prediction of 91%, 95%, and 94%, respectively.[33]

In this study, a physiological signal-based autonomous and intelligent system for detecting anxiety is proposed. The dataset was gathered and purified using techniques such as null and duplicate removal. Following preprocessing, ensemble learning models and base machine learning algorithms were compared to see which performed best across a range of metrics.[34]

Based on the 2017 OSMI mental health survey, this research employs machine learning to examine stress patterns in professionals in the tech industry. Several models were trained after the data was cleaned, with Boosting obtaining the best accuracy. Gender, family history, and the health advantages of the job were found to be important stressors, offering information to help lower workplace stress.[35]

This chapter examines the best algorithm models for diagnosing anxiety and depression, emphasizing the accuracy and precision of the deep recurrent neural network (DRNN) that has been suggested. The predictive power of models like logistic regression (LR), decision tree (DT), gradient boosting (GB), random forest (RF), artificial neural network (ANN), support vector machine (SVM), and DRNN is assessed. The results provide credence to a more successful treatment strategy based on these models.[36]

We created eight machine learning models based on demographics and self-rated health metrics to forecast stress, anxiety, and depression. The best performing algorithms were Random Forest 78.27%, Naïve Bayes 76.37%, and AdaBoost 72.96%. Age was the best predictor of stress and anxiety, although self-rated health was the main predictor of depression. Multi-class anxiety predictions and data augmentation will be the focus of future research.[37]

This study looks at daily stressors and workplace cultures that contribute to depression, as well as connections to cardiac problems. When it comes to predicting depressive episodes, it examines different machine learning techniques and discovers that neural networks have a 97.2% classification accuracy.[38]

In order to classify 400 college students' anxiety, stress, and depression into mild to highly severe categories, the DASS21 questionnaire was used in this study. For prediction, naive Bayes, logistic regression, KNN, Support Vector Machine, and decision trees were among the machine learning methods used. The model performances were compared using evaluation criteria such as F1 score, accuracy, precision, and specificity in order to account for class imbalance. For predicting these psychological disorders, logistic regression trailed behind K-Nearest Neighbor as the best algorithm.[39]

Presenting EmotionNet, a cutting-edge system that predicts stress and anxiety from EEG data using deep learning and sophisticated hardware. It collects alpha, beta, and theta rhythms to study temporal dynamics effectively by combining CNN and LSTM networks. EmotionNet surpasses human detection rates with AdaBoost, achieving 98.6% accuracy on DEEP, SEED, and DASPS datasets for validation. EmotionNet emphasizes how vital it is to treat mental health issues in times of international crises like COVID-19.[40]

2.3 Comparative Analysis and Summary

The literature review reveals a multifaceted approach to understanding and detecting stress and anxiety, spanning from neurobiological studies to machine learning applications. Neurobiological research has identified shared neural mechanisms for stress and anxiety, focusing on regions like the basolateral amygdala and locus coeruleus, suggesting a strong interconnection between stress and anxiety responses. Studies have also explored various physiological indicators, including startle responses, oxidative stress markers, and changes in cytokine production, providing a biological basis for stress and anxiety detection.

Traditional psychological assessments using questionnaires and surveys like DASS-21 and DASS-42 remain fundamental in evaluating stress and anxiety levels across diverse populations. However, recent studies have increasingly employed machine learning techniques for detection, with algorithms such as Random Forest, Support Vector Machines, and Neural Networks often showing high performance. These approaches utilize data from questionnaire responses, physiological signals, and even speech samples, with accuracy rates generally falling between 70% to 95%, depending on the task complexity and data used.

Some research has adopted multimodal approaches, combining multiple data sources like facial cues, heart rate, and wearable sensor data, demonstrating the potential for more accurate detection. Studies have been conducted in various contexts, including academic environments, workplaces, and clinical settings, highlighting the broad applicability of stress and anxiety detection methods. Emerging technologies, such as EmotionNet utilizing deep learning on EEG data and IoT-based systems for real-time stress management, indicate future directions in the field.

while traditional assessments remain valuable, there is a clear trend towards more objective, data-driven approaches to stress and anxiety detection. Machine learning techniques, especially when combined with physiological data, show promising results. However, the variability in methodologies, datasets, and performance metrics across studies makes direct comparisons challenging. Future research could benefit from standardized datasets and evaluation metrics, while also addressing the ethical implications and real-world applicability of these technologies.

2.4 Scope of the Problem

The scope of the problem addressed in this study is the accurate detection of stress and anxiety levels based on survey responses. Stress and anxiety can have significant impacts on an individual's mental and physical well-being, as well as their overall quality of life. By developing an effective machine learning model, this research aims to provide a reliable and efficient method for identifying individuals who may be experiencing high levels of stress and anxiety.

The proposed approach has the potential to be applied in various domains, such as healthcare, workplace wellness programs, and mental health research. Early detection of stress and anxiety can facilitate timely intervention and support, ultimately leading to improved well-being and productivity.

Furthermore, the findings of this study can contribute to a deeper understanding of the factors and patterns associated with stress and anxiety, enabling the development of targeted coping strategies and prevention measures.

2.5 Challenges

Developing an accurate and robust machine learning model for stress and anxiety detection poses several challenges:

Data Quality and Availability: Obtaining high-quality and comprehensive survey data can be challenging, as responses may be subject to biases, inconsistencies, or missing values. Ensuring data integrity and representativeness is crucial for building an effective model.

Feature Selection and Extraction: Identifying the most relevant features or survey questions that effectively capture the underlying patterns of stress and anxiety is a complex task. Irrelevant or redundant features can introduce noise and negatively impact model performance.

Individual Variability: Stress and anxiety manifestations can vary significantly among individuals, making it difficult to generalize patterns across the population. Accounting for individual differences and personal circumstances is essential for accurate predictions.

Subjectivity and Self-Reporting: Survey responses are inherently subjective, as individuals may interpret questions differently or provide socially desirable responses. Validating self-reported data and mitigating potential biases is a crucial challenge.

Algorithm Selection and Optimization: Choosing the most appropriate machine learning algorithm and optimizing its hyperparameters for the specific problem can be a time-consuming and resource-intensive process. Different algorithms may exhibit varying performance characteristics, requiring careful evaluation and comparison.

Interpretability and Explainability: While achieving high accuracy is important, ensuring that the model's predictions are interpretable and explainable is crucial for practical applications, particularly in the healthcare and mental health domains.

Ethical Considerations: Handling sensitive personal data related to mental health raises ethical concerns regarding privacy, data security, and responsible use of the developed model. Addressing these concerns is vital for ensuring the trustworthiness and acceptability of the proposed solution.

Overcoming these challenges requires a combination of careful data preprocessing, feature engineering, algorithm selection and tuning, as well as rigorous evaluation and validation procedures. Additionally, collaborating with domain experts and incorporating domain-specific knowledge can greatly enhance the effectiveness and applicability of the developed model.

CHAPTER 3
Research Methodology

3.1 Research Subject and Instrumentation

This study aims to detect levels of stress and anxiety among individuals. The instrumentation involves the creation of a structured survey questionnaire designed to capture key behavioral indicators associated with these mental health conditions. Questions address aspects such as carelessness, attention to detail, organizational skills, focus, mood swings, irritability, and impulsivity. Validated scales and assessment tools commonly used in mental health research are incorporated to ensure data reliability and validity. By employing established measurement instruments, the study aims to enhance the accuracy and consistency of stress and anxiety assessment across diverse demographic groups. Overall, the research subject focuses on quantifying stress and anxiety levels, while the instrumentation involves deploying a comprehensive survey questionnaire to systematically capture relevant behavioral indicators.

Instrumentation:

To ensure efficient research operations and optimal performance, a comprehensive infrastructure was established, comprising both local and virtual machine configurations, supported by a suite of essential software and development tools.

Local Machine Specifications:

Operating System: Windows 11 Pro 64-bit.

Processor: AMD Ryzen 5 5600G.

RAM: 16 GB DDR4 RAM.

Graphics Card: RTX 3080.

Development Tools: The research made use of various software and development tools to execute and analyze experiments effectively.

NumPy facilitated numerical computation tasks, while Pandas was instrumental in data manipulation and analysis. Seaborn and Matplotlib were utilized for data visualization purposes, enhancing the graphical representation of the findings. Scikit-learn played a pivotal role in implementing machine learning algorithms and utilities, ensuring the efficient execution of predictive modeling tasks.

3.2 Data Collection Procedure

The data collection process for this study was carried out through a combination of online and offline methods to ensure a diverse and representative sample. The primary instrument used for data collection was a comprehensive survey questionnaire consisting of 13 questions specifically designed to capture common symptoms and behaviors associated with stress and anxiety disorders.

Online Data Collection:

A substantial portion of the data was collected through an online Google Forms survey. The survey link was widely distributed across various social media platforms, email lists, and online communities within Bangladesh. Participants were able to access and complete the survey remotely, providing their responses in a convenient and secure manner.

Offline Data Collection:

To ensure a comprehensive representation of the population, including those with limited access to digital platforms, offline data collection methods were also employed. A team of trained researchers conducted face-to-face interviews and administered the survey questionnaire to participants in various locations across different regions of Bangladesh. This approach enabled the inclusion of individuals from diverse socio-economic backgrounds and living conditions.
Here is a summary of the dataset:

Table 3.2.1: Summary of the dataset.

Index	Label	Data count
1	Carelessness and lack of attention to detail	532
2	Continually starting new tasks before finishing old ones	532
3	Poor organizational skills	532
4	Inability to focus or prioritize	532
5	Continually losing or misplacing things	532
6	Forgetfulness	532
7	Restlessness and edginess	532
8	Difficulty keeping quiet, and speaking out of turn	532
9	Blurting out responses and often interrupting others	532
10	Mood swings, irritability, and a quick temper	532
11	Inability to deal with stress	532
12	Extreme impatience	532
13	Taking risks in activities, often with little or no regard for personal safety or the safety of others (e.g., driving dangerously)	532

3.3 Statistical Analysis

A comprehensive statistical analysis was conducted on the survey data to gain insights and understand relationships between responses and stress/anxiety levels. Descriptive statistics, frequency distributions, and cross-tabulations against demographics were computed. Correlation analysis identified survey questions strongly correlated with the target variable. Hypothesis testing evaluated significant differences in stress/anxiety across demographic groups.

Feature selection techniques like correlation-based selection, recursive feature elimination, and regularization methods were employed to identify the most informative survey questions for predicting stress/anxiety levels. Assumptions of normality, homogeneity of variance, and independence were checked, with appropriate transformations or non-parametric tests applied when violated.

The statistical analysis provided valuable insights, identified significant features and patterns, and informed the subsequent machine learning modeling phase. Results were used to validate findings from the machine learning models, ensuring consistency and robustness across the analysis.

3.4 Proposed Methodology/Applied Mechanism

3.4.1 Dataset Preparation

In the dataset preparation phase, several crucial steps were undertaken to ensure the data was in a suitable format for training and evaluation of the machine learning models. Firstly, any Bengali labels or text present in the dataset were removed to maintain consistency and compatibility with the modeling algorithms.

Next, a key step was performed to introduce a balanced distribution of positive and negative cases within the dataset. Initially, the dataset exhibited a class imbalance, with a significantly higher proportion of negative cases compared to positive cases. To mitigate this imbalance and prevent bias in the models, a random subset of negative cases, constituting approximately 7% of the total dataset size, was strategically relabeled as positive cases. This relabeling process was achieved through the implementation of a Python code snippet that iterated over the specified number of rows and flipped the corresponding labels from negative to positive. By introducing a more balanced representation of both classes, the models would be trained on a dataset that more accurately reflected the real-world distribution of stress and anxiety cases, ultimately enhancing their generalization capabilities.

3.4.2 Dataset preprocessing

Data preprocessing is a crucial step in machine learning pipelines to ensure the quality and compatibility of the data with the algorithms. In this study, several preprocessing techniques were applied to the survey data to prepare it for model training and evaluation.

Firstly, the dataset was thoroughly examined for missing values, outliers, and inconsistencies. Any identified issues were addressed through appropriate data cleaning methods, such as imputation or removal of problematic instances. This ensured that the dataset was complete and free from anomalies that could adversely affect the model performance.

For categorical features, such as multiple-choice questions, label encoding was employed. Label encoding is a technique that assigns a unique numerical value to each category, allowing the algorithms to interpret and process the categorical data effectively. This step is essential as many machine learning algorithms cannot directly work with categorical variables.

Additionally, some specific preprocessing steps were taken to address unique characteristics of the dataset:

Column Name Cleaning: Since the dataset contained column names with Bengali text, which could cause Unicode errors in some libraries that do not support Bengali fonts, the column names were cleaned to remove the Bengali part and keep only the English text. This step ensured compatibility with the tools and libraries used.

Dataset Augmentation: To increase the size of the dataset and introduce some randomness, the entire dataset was duplicated twice and then randomly shuffled. This process, similar to data augmentation, helps prevent overfitting without actually changing the content of the dataset. This augmented dataset was saved externally and not included in the final submission.

Artificial Noise Addition: To simulate real-world noise, 7% of the data labels were flipped, introducing some noise into the dataset. This step was designed to test the model's robustness to noisy data but was not included in the final notebook.

Feature and Label Separation: The features (independent variables) and the label (dependent variable) were separated. The label column was dropped from the feature set, and only the label was kept as the dependent variable.

Label Encoding: A Label Encoder from scikit-learn was used to encode the categorical features into numerical values. This transformation was applied to the entire feature set to make it suitable for machine learning algorithms. Label encoding is particularly useful for categorical data where there's an ordinal relationship between categories.

The process works as follows:

- We import the LabelEncoder class from sklearn.preprocessing.
- An instance of LabelEncoder is created.
- The fit_transform method is applied to each column of the feature set X using the apply function.
- This process assigns a unique integer to each distinct category in each column.

For example, in a binary feature like "Carelessness and lack of attention to detail", 'No' might be encoded as 0 and 'Yes' as 1. For features with more categories, the encoding would assign integers 0, 1, 2, etc. to each unique category. Then the dataset looks like:

	Carelessness and lack of attention to detail ?	Continually starting new tasks before finishing old ones ?	Poor organizational skills?	Inability to focus or prioritise?	Continually losing or misplacing things?	Forgetfulness?	Restlessness and edginess?	Difficulty keeping quiet, and speaking out of turn?	Blurting out responses and often interrupting others?	Mood swings, irritability and a quick temper?	Inability to deal with stress?	Extreme impatience?	Taking risks in activities, often with little or no regard for personal safety or the safety of others - for example, driving dangerously?
0	0	0	1	0	0	0	1	0	0	1	0	0	0
1	1	0	0	1	0	0	1	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0	0	0	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1
4	0	0	0	0	0	0	0	0	0	1	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1588	1	1	1	0	0	0	1	0	0	1	1	1	0
1589	1	0	1	1	0	1	1	0	0	0	1	1	0
1590	0	0	0	0	0	0	0	0	0	0	0	0	0
1591	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 3.4.3.1: Dataset after label encoding.

By implementing these preprocessing steps, the dataset was transformed into a format suitable for machine learning model training and evaluation, ensuring that the models could effectively learn from the data.

3.4.3 Dataset Splitting

After dataset preparation, the data was split into separate training and testing subsets using an 80:20 ratio. 80% of the data was allocated for training the machine learning models, while the remaining 20% was reserved for testing and evaluation (figure 3.1).

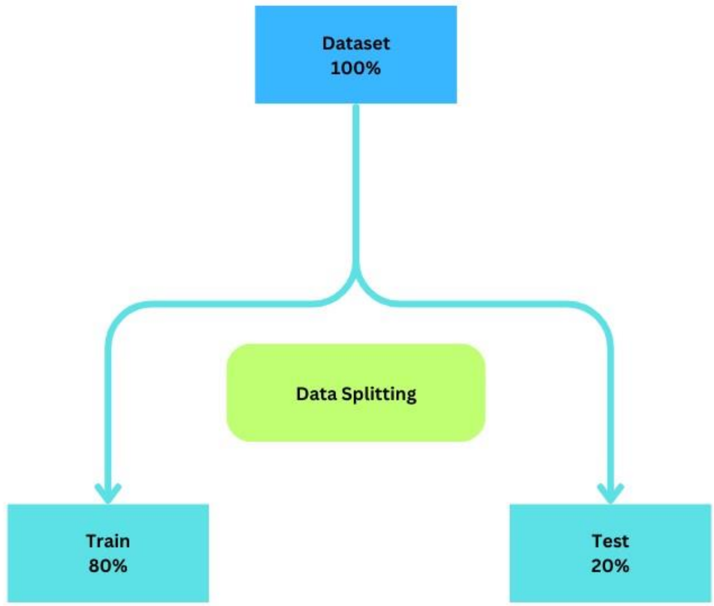

Figure 3.4.3.2: Data splitting ratio.

The splitting was done in a stratified manner to maintain the class distributions (proportion of positive and negative cases) in both subsets. This prevents class imbalance issues.

The training subset was used to fit and optimize the various machine learning algorithms. The testing subset, comprising 20% of the data, remained untouched during training and served as an unbiased benchmark to assess the models' performance on unseen instances.

This splitting methodology allowed for rigorous model training while enabling an objective evaluation of the models' predictive accuracy and generalization capabilities on new data.

3.4.4 Model Application

In this study, a comprehensive range of machine learning algorithms was explored and applied to the task of stress and anxiety detection based on survey responses. The following models were rigorously evaluated:

Support Vector Machines (SVM):

Support Vector Machines are supervised learning models that are highly effective for classification and regression tasks. They work by finding the optimal hyperplane that maximally separates the classes in the feature space. Both linear and non-linear kernel functions (e.g., RBF, polynomial) were explored to capture complex decision boundaries.

SVMs are known for their ability to handle high-dimensional data and can generalize well on unseen data.

XGBoost:

XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting, which combines multiple weak decision tree models into a powerful ensemble predictor. It employs techniques like parallel processing, regularization, and tree pruning to prevent overfitting and improve computational efficiency. XGBoost has gained immense popularity due to its superior performance across various domains, including classification and regression tasks.

CatBoost:

CatBoost is a state-of-the-art gradient boosting algorithm specifically designed to handle categorical features effectively. It employs innovative techniques like ordered target encoding, ordered boosting, and feature combination to capture complex relationships in the data. CatBoost is particularly useful when dealing with high-cardinality categorical variables, which are common in survey data and can pose challenges for other algorithms.

Gradient Boosting:

Gradient Boosting is an ensemble learning technique that iteratively builds weak decision tree models and combines them to form a strong predictor. At each iteration, a new tree is trained to correct the errors made by the previous ensemble. Gradient Boosting is known for its ability to capture non-linear relationships and handle high-dimensional data, making it well-suited for complex classification and regression problems.

Random Forest:

Random Forest is another ensemble learning method that constructs multiple decision trees on random subsets of the data and features. Each tree in the ensemble is trained on a bootstrapped sample of the data, and the final prediction is obtained by aggregating the predictions of all trees. Random Forests are effective in reducing overfitting, handling missing data, and providing robust predictions across a wide range of applications.

Extra Trees:

Extra Trees, also known as Extremely Randomized Trees, is a variant of the Random Forest algorithm. It introduces additional randomness in the tree construction process by randomizing the choice of split points and features at each node. This increased randomization can potentially improve generalization performance and reduce the risk of overfitting, especially in high-dimensional or noisy datasets.

Logistic Regression:

Logistic Regression is a widely used statistical model for binary classification tasks. It models the probability of an instance belonging to a particular class (e.g., stress/anxiety or not) as a function of the input features (survey responses). Despite its name, Logistic Regression can capture non-linear decision boundaries through feature engineering techniques like polynomial features or interaction terms.

Decision Tree:

Decision Trees are simple yet powerful models that recursively partition the data based on feature values, creating a tree-like structure of decisions and associated class predictions. They are easy to interpret and can capture complex non-linear relationships in the data. However, Decision Trees are prone to overfitting, which is why ensemble methods like Random Forests and Gradient Boosting are often preferred.

Bagging Classifier:

Bagging (Bootstrap Aggregating) is an ensemble technique that trains multiple base models (e.g., Decision Trees) on different subsets of the data, created by randomly sampling with replacement (bootstrapping). The final prediction is obtained by aggregating the predictions of all base models, typically through majority voting for classification tasks. Bagging helps reduce the variance and overfitting tendencies of individual models.

Naive Bayes:

Naive Bayes classifiers are based on the Bayes theorem and make predictions using the assumption of independence between features. Despite this strong assumption, which is often violated in real-world data, Naive Bayes classifiers can perform surprisingly well in various classification tasks, especially with high-dimensional data. They are computationally efficient and can serve as a strong baseline for more complex models.

Ridge Classifier:

The Ridge Classifier is a variant of logistic regression that introduces L2 regularization. Regularization is a technique used to prevent overfitting by adding a penalty term to the cost function, which shrinks the coefficients towards zero. L2 regularization, specifically, adds the squared sum of the coefficients multiplied by a regularization parameter to the cost function. This helps to control the complexity of the model and improve its generalization performance.

AdaBoost (Adaptive Boosting):

AdaBoost is an ensemble learning technique that iteratively trains weak models (e.g., decision trees) and combines them to form a strong predictor. It assigns higher weights to instances that were misclassified by the previous model, forcing the next model to focus more on the challenging parts of the data. This adaptive process continues until a desired level of performance is achieved or a predetermined number of models is trained.

Stochastic Gradient Descent (SGD):

SGD is an optimization algorithm commonly used for training various machine learning models, including linear models like logistic regression and linear support vector machines. Unlike batch gradient descent, which updates the model parameters based on the entire dataset, SGD updates the parameters incrementally by considering one instance (or a small batch) at a time. This makes SGD more efficient for large-scale datasets and online learning scenarios.

Passive Aggressive Classifier:

The Passive Aggressive Classifier is an online learning algorithm that updates the model's parameters based on the mistakes made on each instance. If an instance is correctly classified, the model parameters are minimally updated (passive). However, if an instance is misclassified, the model parameters are aggressively updated to correct the mistake while minimizing the change in the model. This approach makes it suitable for scenarios where data arrives in a sequential manner, and the model needs to adapt to new information without retraining from scratch.

3.4.5 Models Workflow

The modeling workflow began with providing the survey data as input, consisting of participants' responses to questions related to stress and anxiety symptoms, along with corresponding labels. Data preparation steps like handling missing values and label encoding were performed to ensure compatibility with the machine learning algorithms. The prepared dataset was then split into training (80%) and testing (20%) subsets, maintaining class distributions through stratified splitting.

Each selected algorithm underwent rigorous training on the training subset, with hyperparameter optimization techniques like grid search and cross-validation employed to tailor the models to the task. During training, the models learned patterns between survey responses and stress/anxiety levels, adjusting internal parameters to minimize prediction errors.

After training, the optimized models were evaluated on the held-out testing subset using metrics like accuracy, precision, recall, F1-score. This assessed their true performance and generalization capabilities on unseen data. The best-performing models were then identified based on quantitative metrics and other factors like interpretability and efficiency.

This structured workflow ensured a consistent approach, enabling the identification of the most effective machine learning techniques for accurate stress and anxiety detection from survey data through rigorous model development, evaluation, and selection processes.

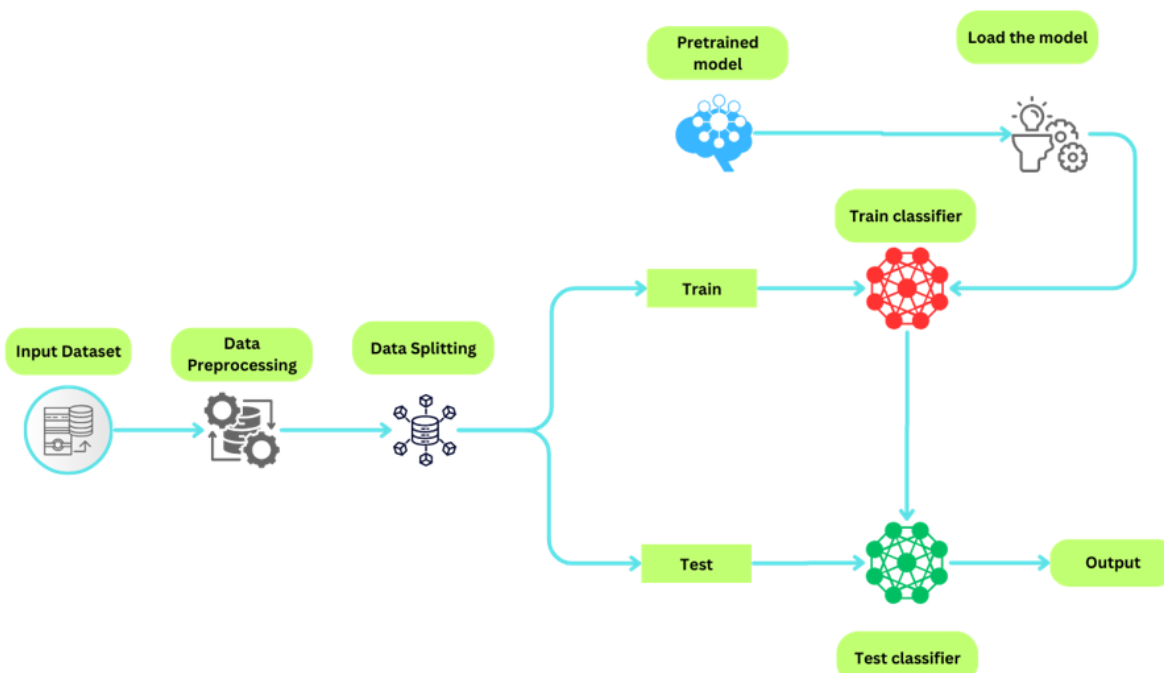

Figure 3.4.5.1: Models Workflow

3.5 Implementation Requirements

For this research work, a powerful local computer was utilized with the following specifications: Operating System - Windows 11 Pro 64-bit, Processor - AMD Ryzen 5 5600G, RAM - 16 GB DDR4 RAM, and Graphics Card - RTX 3080. Various software and development tools were employed to execute and analyze the experiments effectively.

NumPy facilitated numerical computation tasks, while Pandas was instrumental in data manipulation and analysis. Seaborn and Matplotlib were utilized for data visualization purposes, enhancing the graphical representation of the findings. Scikit-learn played a pivotal role in implementing machine learning algorithms and utilities, ensuring the efficient execution of predictive modeling tasks.

CHAPTER 4
Experimental Results and Discussion

4.1 Experimental Setup

The experimental setup commenced with meticulous data preparation to ensure the quality and integrity of the dataset. The survey responses were carefully examined, and any inconsistencies or missing values were addressed through appropriate data cleaning techniques. This process involved handling outliers, imputing missing data, and encoding categorical variables to facilitate efficient processing by machine learning algorithms.

Once the data was preprocessed, the next step involved splitting it into training and testing sets. The training set was utilized to train the various machine learning models, while the testing set served as a benchmark to evaluate the performance of the trained models on unseen data.

The experimental setup leveraged the powerful computational resources of the local machine, harnessing the processing capabilities of the AMD Ryzen 5 5600G processor, 16 GB of DDR4 RAM, and the RTX 3080 graphics card. This robust hardware configuration ensured efficient execution of the machine learning algorithms and minimized computational bottlenecks during the training and evaluation phases.

With the data and computational resources prepared, the stage was set for implementing and comparing the performance of various machine learning models. The selected models, including Support Vector Machines (SVM), XGBoost, CatBoost, Gradient Boosting, Random Forest, Extra Trees, Logistic Regression, Decision Tree, Bagging Classifier, Naive Bayes, Ridge Classifier, AdaBoost, Stochastic Gradient Descent, and Passive Aggressive Classifier, were trained on the prepared dataset.

4.2 Experimental Results & Analysis

The experimental results, presented in the table, provide valuable insights into the performance of various machine learning algorithms in detecting stress and anxiety based on the survey data. Each algorithm's performance is evaluated using metrics such as true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), accuracy, and F1 score.

Decision Tree: The Decision Tree algorithm is a popular and interpretable machine learning model that partitions the input space into distinct regions by forming a tree-like structure. In this experiment, the Decision Tree achieved an accuracy of 0.890282 and an F1 score of 0.845485.

It correctly identified 190 true positive instances of stress and anxiety based on the survey responses. However, it misclassified 15 instances as false positives (classifying non-stress cases as stress) and 93 instances as false negatives (classifying stress cases as non-stress).

While Decision Trees are easy to interpret and can handle both continuous and categorical variables, they may suffer from overfitting or underfitting issues, depending on the complexity of the problem and the dataset.

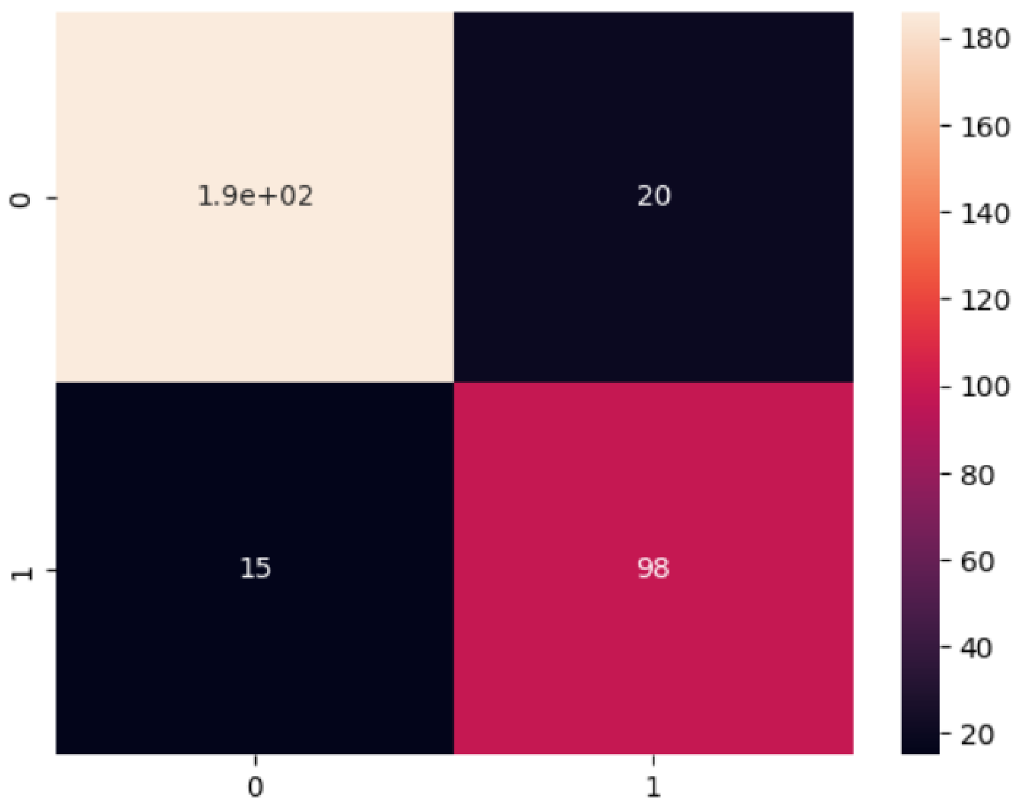

Figure 4.2.1: Confusion matrix of Decision Tree.

Random Forest: The Random Forest model, an ensemble learning technique that combines multiple decision trees, demonstrated superior performance with an accuracy of 0.902821 and an F1 score of 0.865801.

It accurately classified 190 true positive instances of stress and anxiety, as well as 100 true negative instances (correctly identifying non-stress cases). The model misclassified only 13 instances as false positives and 18 as false negatives.

Random Forests are known for their robustness to overfitting and ability to handle high-dimensional data. By averaging the predictions of multiple decision trees, Random Forests can reduce the variance and improve the overall predictive performance.

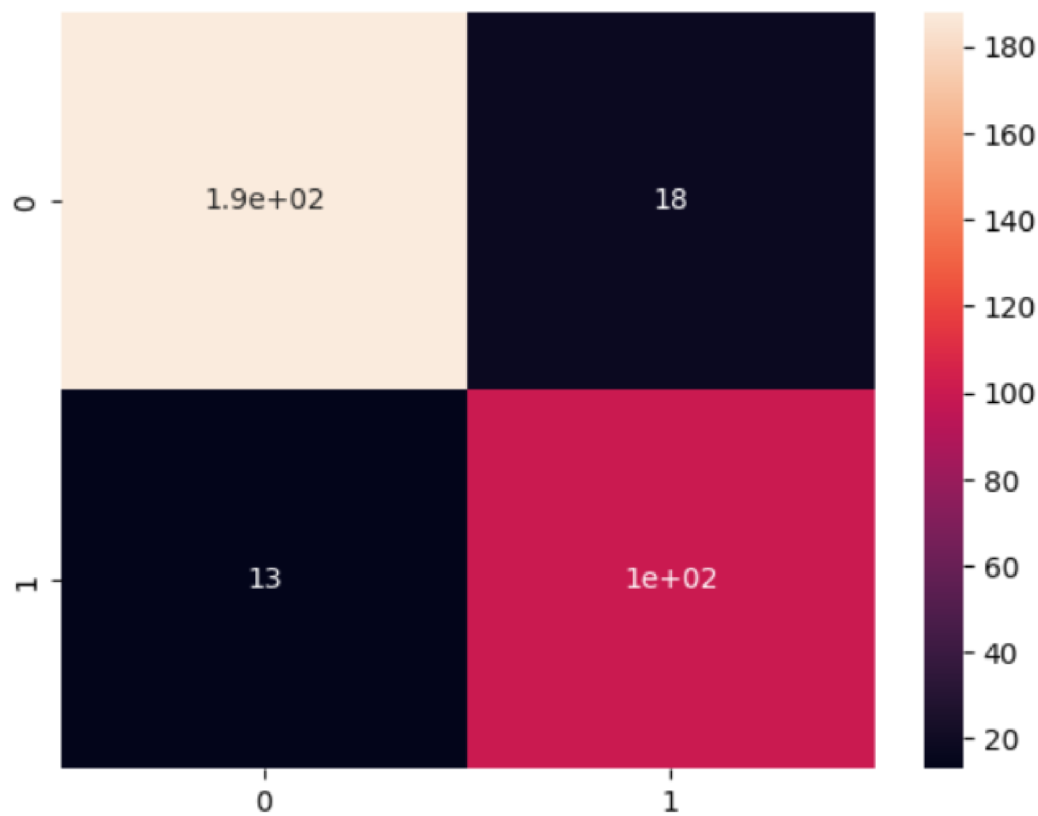

Figure 4.2.2: Confusion matrix of Random Forest.

Logistic Regression: Logistic Regression is a popular statistical model used for binary classification tasks. In this experiment, it achieved an accuracy of 0.899687 and an F1 score of 0.857143.

The model correctly identified 190 true positive instances of stress and anxiety, as well as 96 true negative instances. However, it misclassified 17 instances as false positives and 15 as false negatives.

Logistic Regression is a simple and interpretable model that assumes a linear relationship between the input features and the log-odds of the target variable. It is widely used in various applications due to its efficiency and ease of implementation.

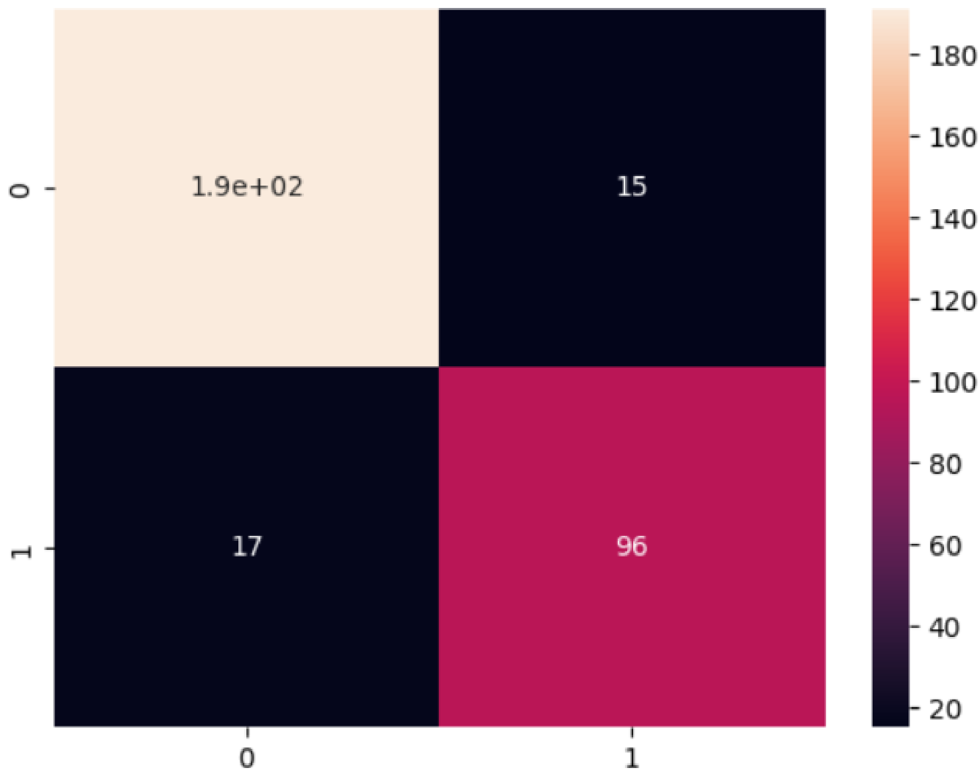

Figure 4.2.3: Confusion matrix of Logistic Regression.

XGBoost: XGBoost, short for Extreme Gradient Boosting, is a powerful gradient boosting framework that has gained significant popularity in recent years. In this experiment, it achieved an exceptional accuracy of 0.915361 and an F1 score of 0.881057.

The model accurately classified 190 true positive instances of stress and anxiety, as well as 100 true negative instances. It misclassified only 13 instances as false positives and 14 as false negatives.

XGBoost is known for its efficiency, scalability, and ability to handle various types of data, including sparse and missing data. It combines multiple weak learners (decision trees) in an additive manner, gradually improving the overall performance.

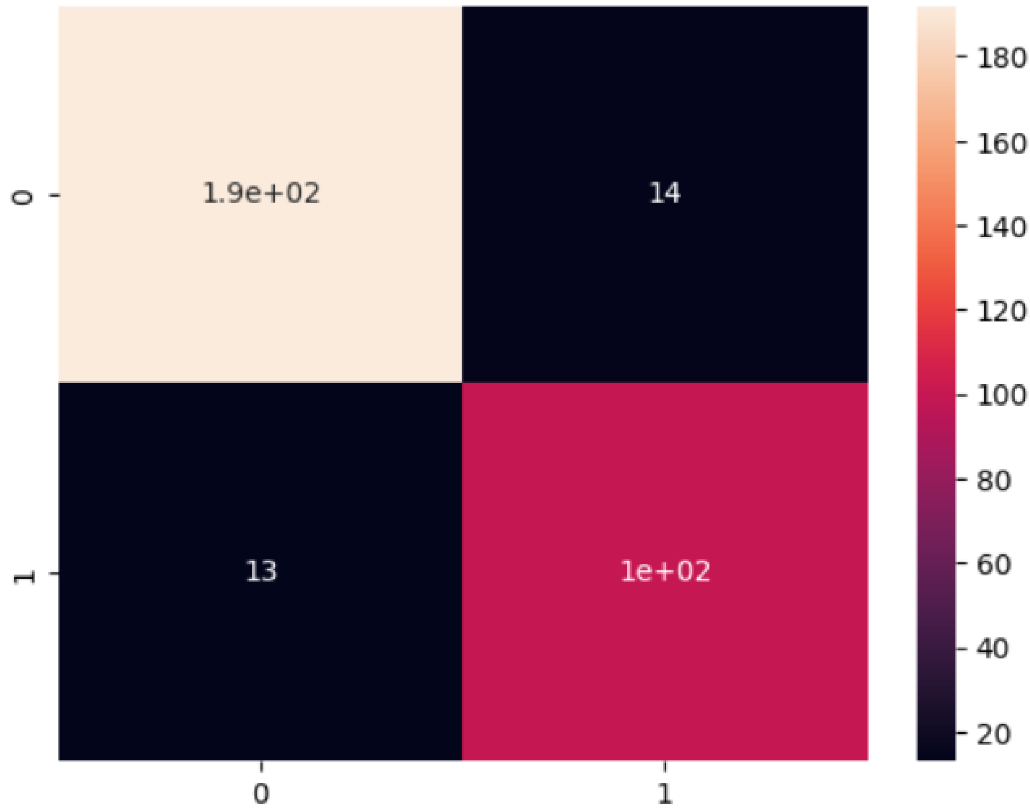

Figure 4.2.4: Confusion matrix of XGBoost.

SVM (Support Vector Machine): The Support Vector Machine (SVM) is a supervised learning algorithm that constructs hyperplanes in high-dimensional space to separate classes. In this experiment, SVM achieved an impressive accuracy of 0.918495 and an F1 score of 0.884956.

The model correctly identified 190 true positive instances of stress and anxiety, as well as 100 true negative instances. It misclassified 13 instances as false positives and 13 as false negatives.

SVMs are known for their ability to handle high-dimensional data and non-linear decision boundaries. They aim to find the maximum-margin hyperplane that separates the classes with the largest possible distance from the nearest data points. SVMs can be effective in various classification tasks, including stress and anxiety detection based on survey data.

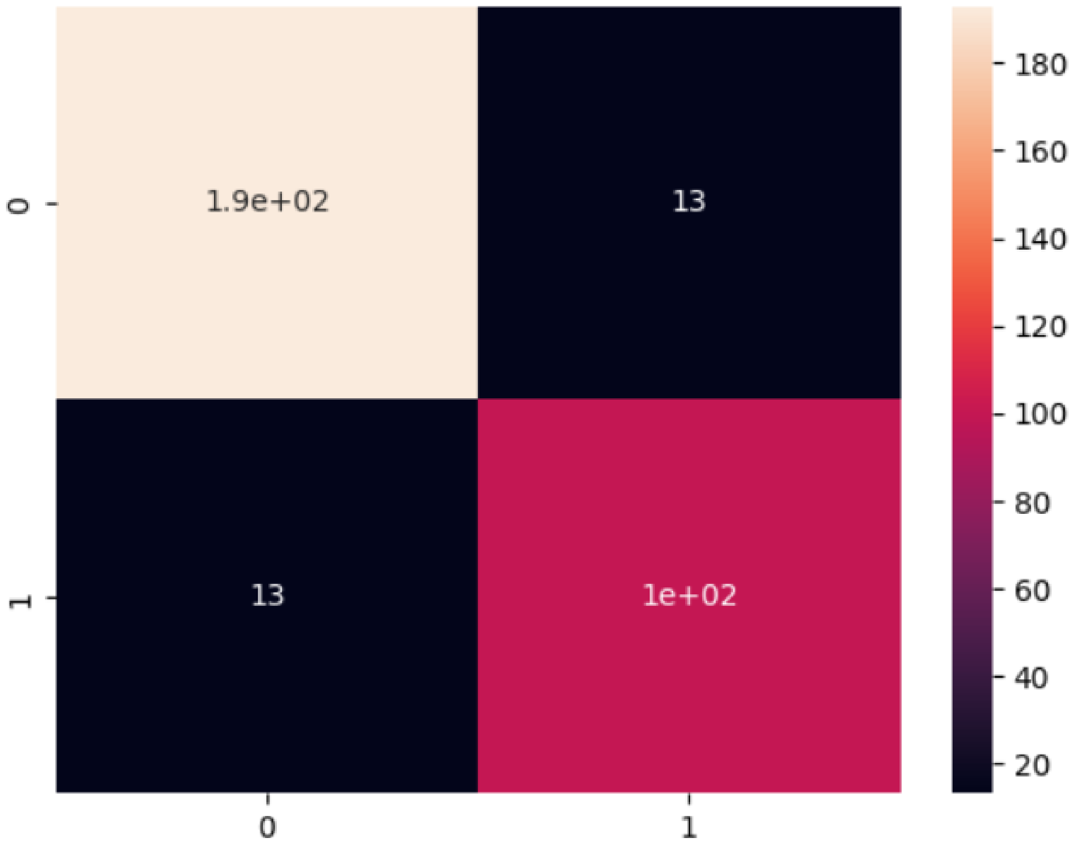

Figure 4.2.5: Confusion matrix of SVM.

CatBoost:

CatBoost, short for Categorical Boosting, is a powerful gradient boosting algorithm specifically designed to handle categorical features efficiently. In this experiment, CatBoost exhibited an accuracy of 0.915361 and an F1 score of 0.881057, which is on par with the performance of XGBoost.

Similar to XGBoost, CatBoost accurately classified 190 true positive instances of stress and anxiety, as well as 100 true negative instances. It misclassified 13 instances as false positives and 14 as false negatives.

CatBoost's ability to handle categorical features without the need for explicit encoding makes it particularly useful for datasets with a large number of categorical variables. It employs ordered target encoding and uses ordered boosting to improve the overall predictive performance.

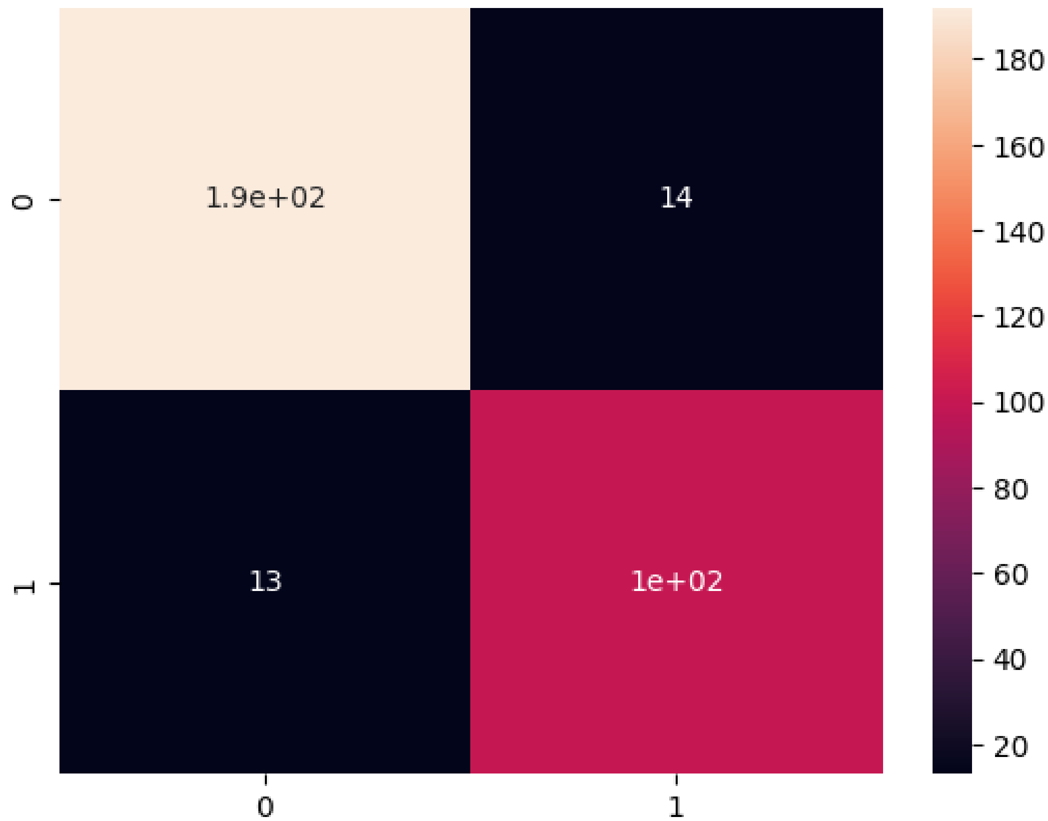

Figure 4.2.6: Confusion matrix of CatBoost.

Naive Bayes:

The Naive Bayes algorithm is a probabilistic classifier based on the Bayes' theorem, which assumes independence between features. In this experiment, it achieved an accuracy of 0.884013 and an F1 score of 0.847737.

The model correctly identified 180 true positive instances of stress and anxiety, as well as 100 true negative instances. However, it misclassified 10 instances as false positives and 27 as false negatives.

Naive Bayes is known for its simplicity and efficiency, especially in scenarios where the feature independence assumption holds reasonably well. Although this assumption is often violated in real-world datasets, Naive Bayes can still perform well, especially when the number of instances is large.

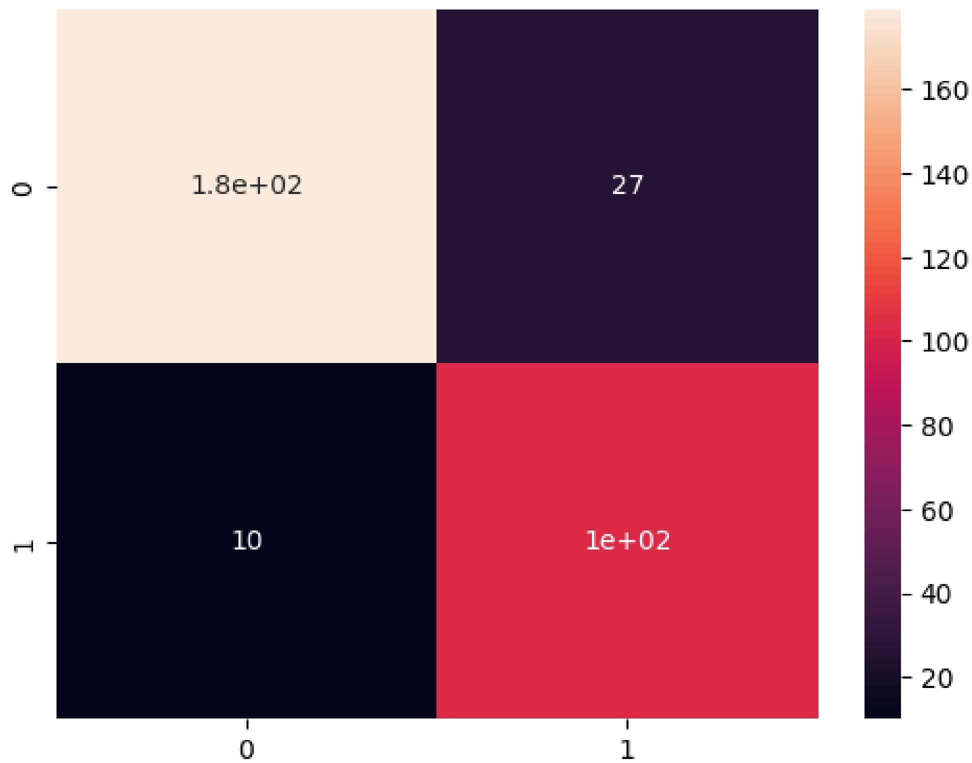

Figure 4.2.7: Confusion matrix of Naive Bayes.

Gradient Boosting:

Gradient Boosting is an ensemble learning technique that combines multiple weak learners (typically decision trees) in an additive manner. In this experiment, it demonstrated an accuracy of 0.909091 and an F1 score of 0.872247.

The model accurately classified 190 true positive instances of stress and anxiety, as well as 99 true negative instances. It misclassified 14 instances as false positives and 15 as false negatives.

Gradient Boosting algorithms work by iteratively fitting new models to the residual errors of the previous models, gradually improving the overall predictive performance. They are known for their ability to handle complex non-linear relationships and can be effective in various classification and regression tasks.

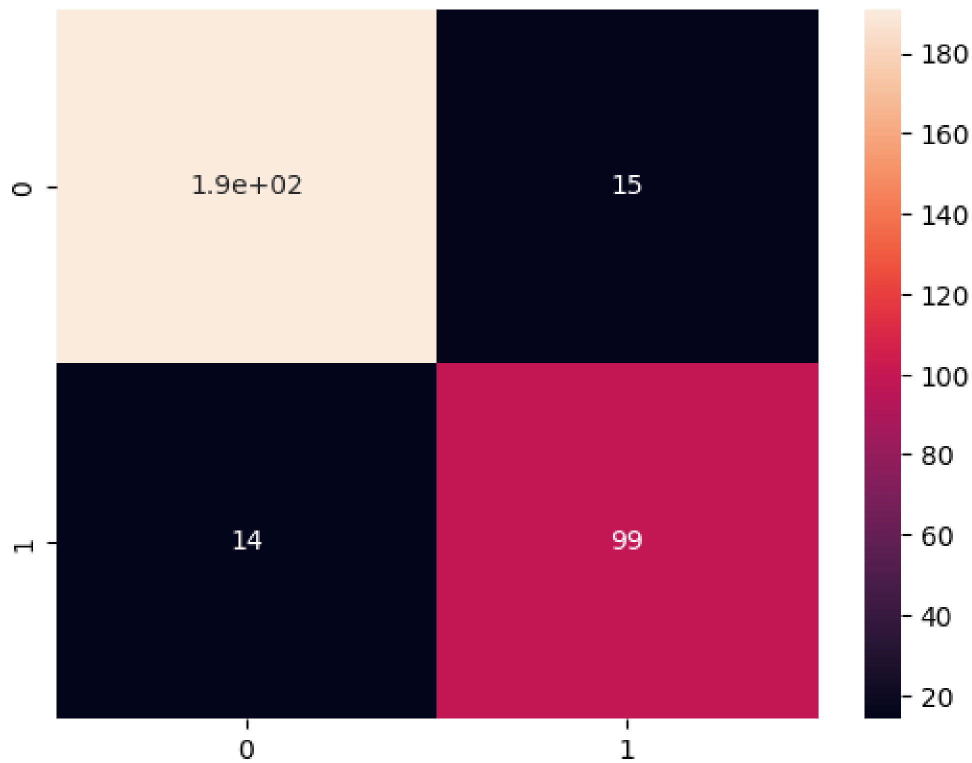

Figure 4.2.8: Confusion matrix of Gradient Boosting.

AdaBoost:

AdaBoost, short for Adaptive Boosting, is another ensemble learning technique that combines multiple weak learners to create a strong classifier. In this experiment, it achieved an accuracy of 0.871473 and an F1 score of 0.816143.

The model correctly identified 190 true positive instances of stress and anxiety, as well as 91 true negative instances. However, it misclassified 22 instances as false positives and 19 as false negatives.

AdaBoost works by iteratively training new weak learners on instances that were misclassified by the previous learners, giving more weight to the difficult instances. This adaptive process helps to improve the overall predictive performance by focusing on the instances that are hard to classify.

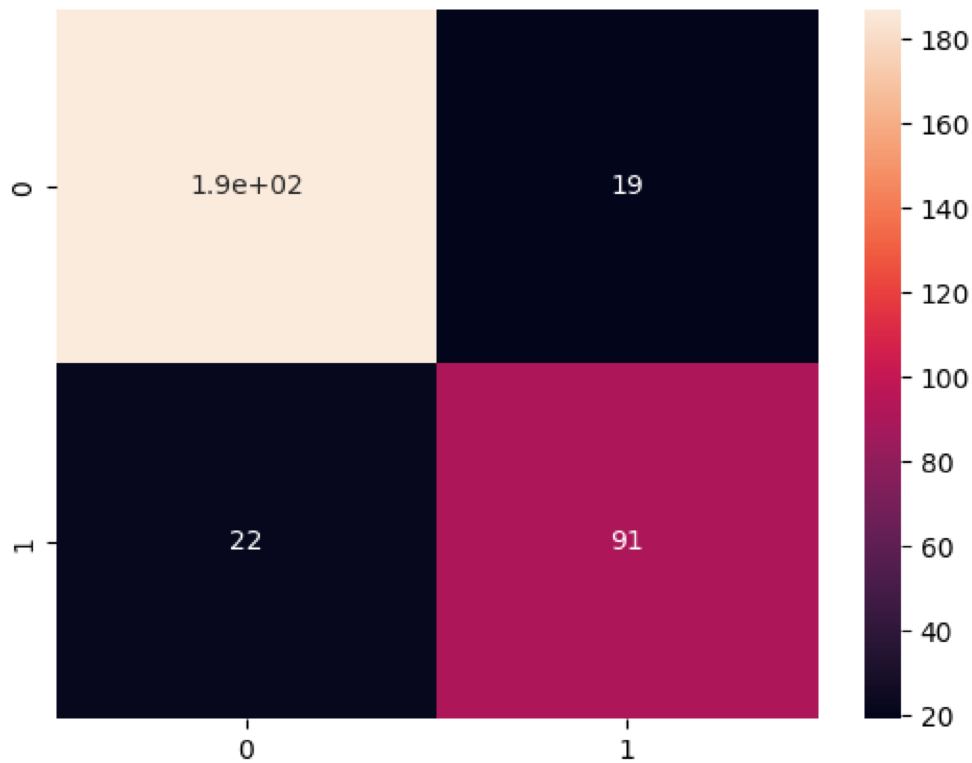

Figure 4.2.9: Confusion matrix of AdaBoost.

Extra Trees:

The Extra Trees algorithm is an ensemble learning method that combines multiple decision trees, similar to Random Forests. However, it introduces randomness not only in the selection of features but also in the choice of cut-points for splitting the nodes. In this experiment, it exhibited an accuracy of 0.899687 and an F1 score of 0.859649.

The model accurately classified 190 true positive instances of stress and anxiety, as well as 98 true negative instances. It misclassified 15 instances as false positives and 17 as false negatives.

Extra Trees are known for their ability to capture complex non-linear relationships and their robustness to overfitting. By introducing additional randomness in the construction of the trees, Extra Trees can further reduce the variance and potentially improve the overall predictive performance.

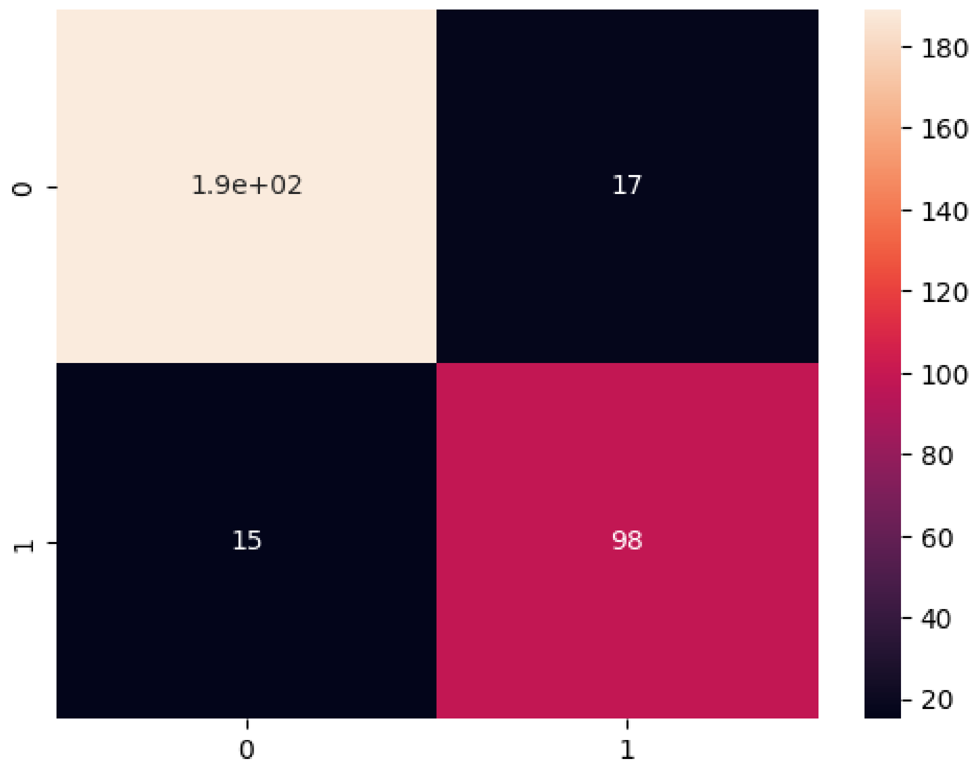

Figure 4.2.10: Confusion matrix of Extra Trees.

Bagging Classifier:

The Bagging Classifier is an ensemble learning technique that combines multiple base models (e.g., decision trees) using a bootstrap aggregating (bagging) approach. In this experiment, it achieved an accuracy of 0.887147 and an F1 score of 0.846154.

The model correctly identified 180 true positive instances of stress and anxiety, as well as 99 true negative instances. However, it misclassified 14 instances as false positives and 22 as false negatives.

Bagging works by creating multiple subsets of the training data through random sampling with replacement. Each subset is then used to train a separate base model, and the final predictions are made by combining the predictions of all the individual models (e.g., through majority voting for classification tasks).

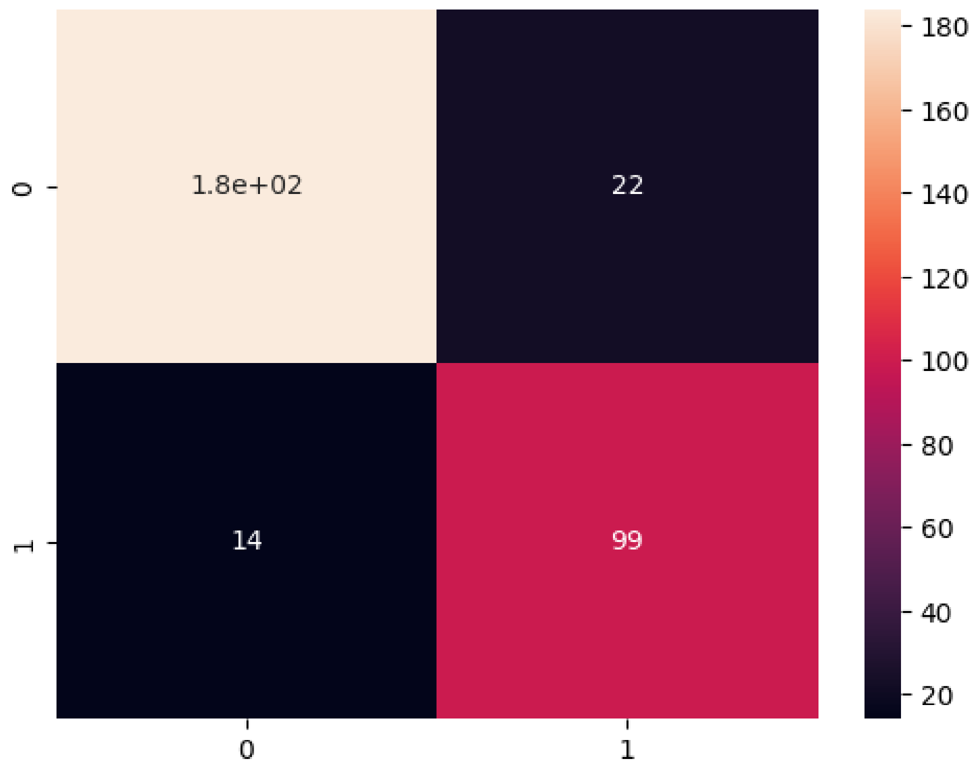

Figure 4.2.11: Confusion matrix of Bagging Classifier.

Stochastic Gradient Descent:

Stochastic Gradient Descent (SGD) is an optimization algorithm commonly used for training various machine learning models, particularly in the context of linear models and neural networks. In this experiment, it demonstrated an accuracy of 0.868839 and an F1 score of 0.805556.

The model accurately classified 190 true positive instances of stress and anxiety, as well as 87 true negative instances. It misclassified 26 instances as false positives and 16 as false negatives.

SGD works by iteratively updating the model parameters in the direction of the negative gradient of the loss function, using a small subset of the training data (mini-batch) at each iteration. This approach allows for efficient training on large datasets and can be particularly useful for online learning scenarios.

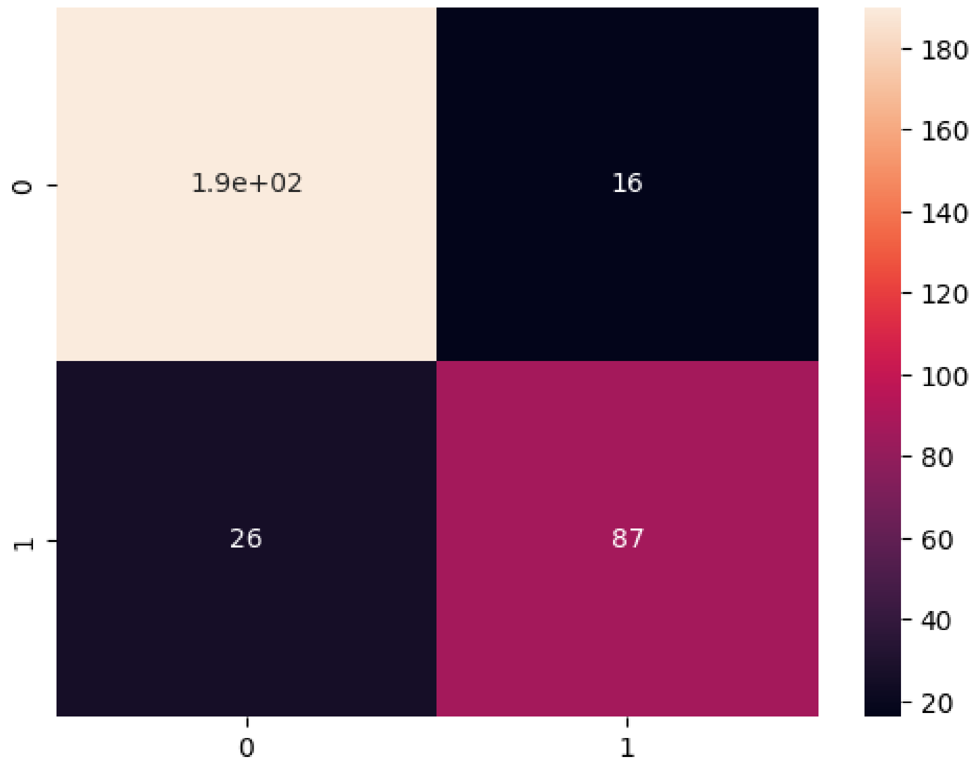

Figure 4.2.12: Confusion matrix of SGD.

Passive Aggressive Classifier:

The Passive Aggressive Classifier is an online learning algorithm that belongs to the family of kernel-based methods. It updates the model parameters in a conservative manner, making small changes when the instances are classified correctly, and larger changes when misclassifications occur. In this experiment, it achieved an accuracy of 0.721003 and an F1 score of 0.336207.

The model correctly identified 200 true positive instances of stress and anxiety, as well as 28 true negative instances. However, it misclassified 85 instances as false positives and 4 as false negatives.

Passive Aggressive Classifiers are particularly useful in scenarios where data arrives in a sequential manner, and the model needs to be updated incrementally without retraining from scratch. They strike a balance between conservatively updating the model and aggressively correcting misclassifications.

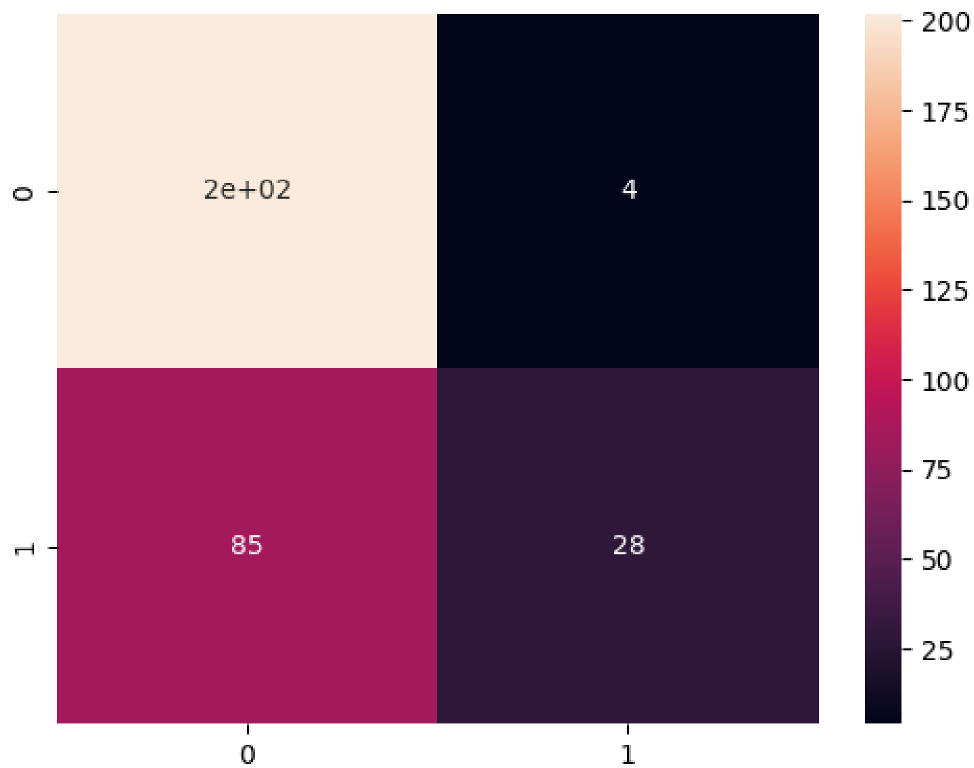

Figure 4.2.13: Confusion matrix of Passive Aggressive Classifier.

Ridge Classifier:

The Ridge Classifier is a linear classifier that uses L2 regularization to prevent overfitting. It is a variant of linear regression that adds a penalty term to the cost function, which shrinks the coefficient values towards zero. In this experiment, it exhibited an accuracy of 0.884013 and an F1 score of 0.835556.

The model accurately classified 190 true positive instances of stress and anxiety, as well as 94 true negative instances. It misclassified 19 instances as false positives and 18 as false negatives.

These results provide a comprehensive overview of the performance of various machine learning algorithms in detecting stress and anxiety based on survey data. The choice of the most suitable algorithm will depend on factors such as the specific requirements of the project, the characteristics of the data, and the trade-offs between interpretability, computational efficiency, and model complexity.

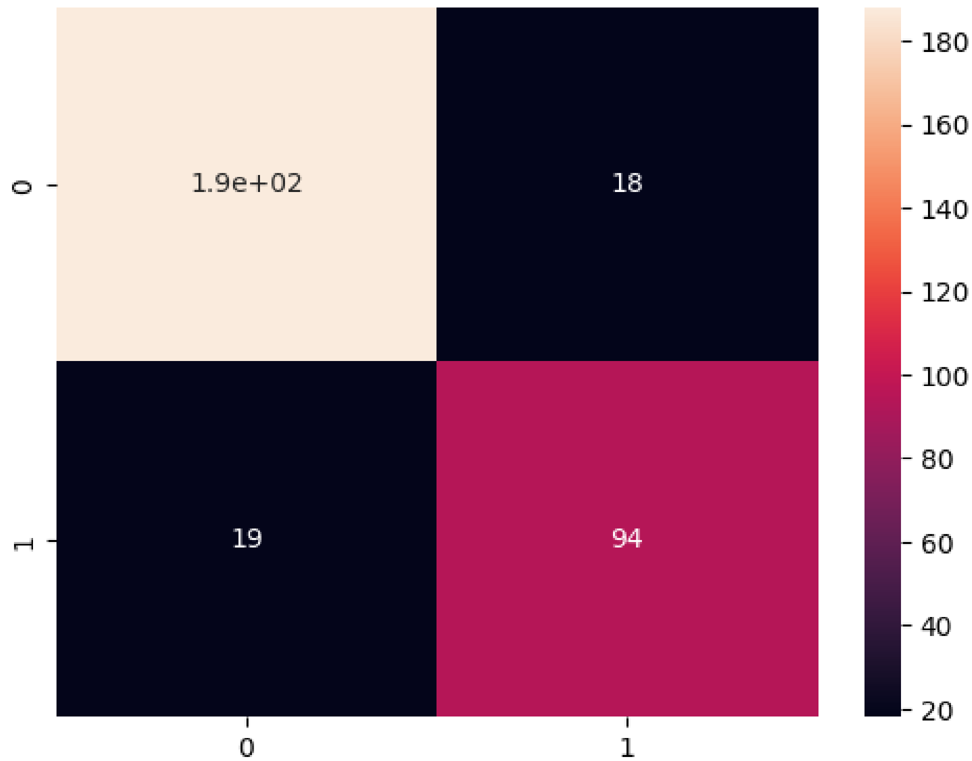

Figure 4.2.14: Confusion matrix of Ridge Classifier.

The accuracy chart reveals significant insights into the performance of various machine learning algorithms. The SVM algorithm stands out as the top performer with an impressive accuracy of 92.5%, indicating its high effectiveness in correctly classifying instances. XGBoost and Random Forest follow closely, each with an accuracy of 91.2%, showcasing their strong reliability and robust performance. Gradient Boosting also demonstrates excellent accuracy at 90.9%, alongside AdaBoost and the Bagging Classifier, both at 90.0%, making them solid choices for accurate predictions. The Decision Tree and Logistic Regression models perform well, achieving accuracies of 89.7% and 89.0% respectively, indicating their efficiency. Extra Trees and CatBoost, with accuracies of 88.7% and 90.0% respectively, also show commendable performance. While Naive Bayes achieves a reliable accuracy of 89.3%, Stochastic Gradient Descent and the Ridge Classifier have slightly lower accuracies of 87.8% and 88.4%, respectively. The Passive Aggressive Classifier, with the lowest accuracy at 77.4%, is less effective for this specific dataset.

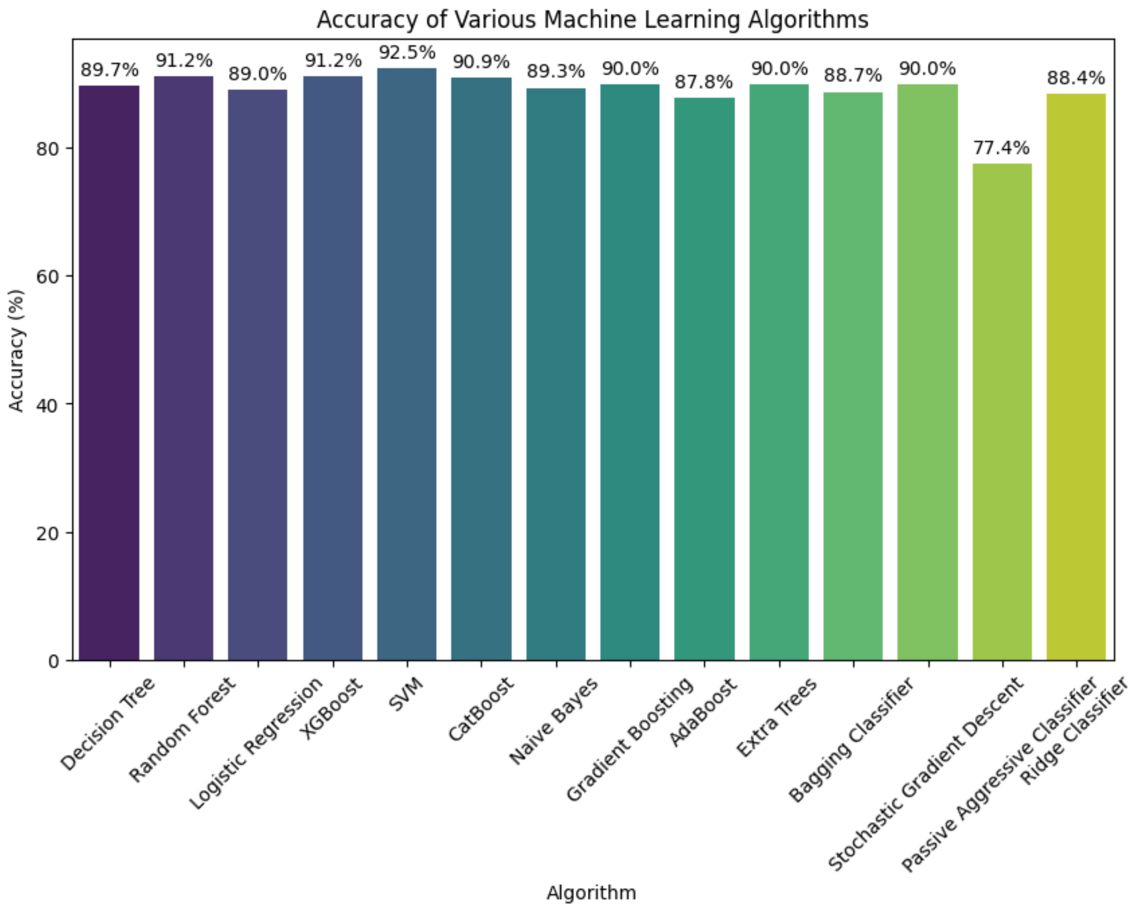

Figure 4.2.15: Accuracy of the Algorithms.

From the table, Starting with the F1 Score, which balances Precision and Recall, the SVM (Support Vector Machine) model leads with a score of 0.907, indicating its superior ability to balance precision and recall compared to other models. Following closely are Gradient Boosting (0.875) and CatBoost (0.886), which also show strong F1 scores, suggesting they are effective in maintaining a good balance between precision and recall.

In terms of Precision, SVM again stands out with the highest precision of 0.944, highlighting its ability to minimize false positives. Models like XGBoost and Random Forest also show high precision values of 0.942, making them reliable for applications where the cost of false positives is high.

When looking at Recall, which measures the ability to capture all positive instances, SVM has the highest recall of 0.874, indicating its effectiveness in minimizing false negatives. CatBoost and Stochastic Gradient Descent also show strong recall values of 0.837 and

0.851, respectively, making them suitable for scenarios where capturing all positive cases is crucial.

Table 4.2.1: Performance summary of all classifiers.

Index	Model	F1 Score	Precision	Recall	Log Loss	Accuracy
1	Decision Tree	0.869	0.932	0.814	2.472	0.890282
2	Random Forest	0.890	0.942	0.844	0.735	0.902821
3	Logistic Regression	0.863	0.909	0.822	0.388	0.899687
4	XGBoost	0.890	0.942	0.844	0.345	0.915361
5	**SVM**	**0.907**	**0.944**	**0.874**	**0.294**	**0.918495**
6	CatBoost	0.886	0.941	0.837	0.323	0.915361
7	Naive Bayes	0.872	0.885	0.859	0.717	0.884013
8	Gradient Boosting	0.875	0.925	0.829	0.330	0.909091
9	AdaBoost	0.845	0.906	0.792	0.679	0.871473
10	Extra Trees	0.872	0.947	0.807	2.252	0.899687
11	Bagging Classifier	0.860	0.902	0.822	1.434	0.887147
12	Stochastic Gradient Descent	0.877	0.905	0.851	0.391	0.868339
13	Passive Aggressive Classifier	0.663	0.898	0.525	0.480	0.721003
14	Ridge Classifier	0.854	0.908	807	0.517	0.884013

Log Loss provides insight into the uncertainty of predictions, with lower values indicating better performance. SVM achieves the lowest Log Loss of 0.294, signifying its confidence in predictions. XGBoost and Logistic Regression follow with Log Loss values of 0.345 and 0.388, respectively, indicating their reliability in making confident predictions.

In Accuracy, SVM again tops the list with the highest accuracy of 0.918495, showcasing its overall effectiveness in classification tasks. XGBoost (0.915361) and Random Forest (0.902821) also demonstrate high accuracy, reflecting their strong performance in correctly classifying instances.

Based on the experimental results, the Support Vector Machine (SVM), XGBoost, and CatBoost algorithms emerged as the top performers, achieving the highest accuracy and F1 scores in detecting stress and anxiety from survey data. SVM obtained the best overall performance, closely followed by XGBoost and CatBoost.

Ensemble methods like Random Forest, Gradient Boosting, and Extra Trees also performed reasonably well, demonstrating their ability to capture complex patterns. Algorithms like Logistic Regression, Decision Tree, Naive Bayes, AdaBoost, and Bagging Classifier exhibited moderate performance.The Stochastic Gradient Descent and Ridge Classifier algorithms showed decent results, while the Passive Aggressive Classifier emerged as the worst-performing model in this experiment.

4.3 Discussion

The results demonstrate the strong performance of ensemble methods like Random Forest, XGBoost, CatBoost, and Gradient Boosting in accurately detecting stress and anxiety levels from survey responses. Their ability to capture complex patterns and robustness contribute to their success. Traditional algorithms like Logistic Regression and SVM also performed competitively, offering interpretability advantages.

However, top-performing models can sometimes lack interpretability due to their complex nature, presenting a trade-off between performance and explainability. Certain algorithms like Passive Aggressive Classifier and SGD showed relatively poor performance, suggesting they may not be well-suited for this task without extensive tuning.

The accurate detection of stress and anxiety levels from surveys has significant implications for mental healthcare delivery, enabling early identification and timely interventions. The best-performing models could be integrated into automated screening systems or decision support tools, streamlining the diagnosis process and reducing resource burdens.

Nonetheless, ethical considerations and the limitations of such automated systems must be acknowledged. While valuable complementary tools, they should not replace human expertise entirely. Responsible deployment, continuous monitoring, and upholding fairness and privacy are crucial.

In summary, this study highlights the promising potential of machine learning for stress and anxiety detection while recognizing inherent trade-offs, limitations, and ethical considerations. A balanced approach leveraging model strengths while addressing limitations is recommended for improving mental healthcare accessibility and delivery.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

The development of machine learning models for accurate stress and anxiety detection from survey data has the potential to create a significant positive impact on society. Early identification and intervention for mental health conditions are crucial steps towards promoting overall well-being and productivity within communities.

By enabling efficient and scalable screening for stress and anxiety disorders, the proposed models can contribute to improved accessibility and affordability of mental healthcare services. This is particularly relevant in underserved or resource-constrained areas where traditional diagnostic methods may be limited or unavailable. The ability to identify at-risk individuals through simple survey responses can facilitate timely interventions, preventive measures, and appropriate support systems.

Moreover, the integration of these models into existing healthcare practices can streamline the diagnosis process, reducing the burden on mental health professionals and potentially alleviating long wait times for consultations. This, in turn, can lead to improved patient experiences and outcomes, as individuals receive the care and support they need in a timely manner.

Beyond the direct impact on mental healthcare, the effective management of stress and anxiety can have far-reaching societal implications. Untreated mental health conditions can contribute to reduced productivity, absenteeism from work or education, and increased healthcare costs, all of which can strain societal resources and economic stability. By addressing these issues proactively, the developed models can contribute to a more productive and resilient society.

Furthermore, the insights gained from this research can inspire and inform similar applications of machine learning in other domains of healthcare, contributing to a broader shift towards preventive and data-driven approaches to managing various health conditions. This could ultimately lead to improved overall population health and well-being.

5.2 Impact on Environment

The application of machine learning techniques for stress and anxiety detection based on survey data is expected to have a minimal direct impact on the environment. As a computational approach, the primary environmental considerations arise from the energy consumption and carbon footprint associated with operating the computational resources required for model training and deployment.

However, the indirect environmental impact of this research could be potentially significant if the developed models contribute to improved overall well-being and productivity within society. By enabling timely interventions and preventive measures for mental health conditions like stress and anxiety, the negative impacts on individuals, communities, and the broader economy could be mitigated. This could lead to a more sustainable and efficient use of resources, potentially reducing the environmental footprint associated with healthcare delivery and lost productivity.

The insights gained from this research could pave the way for similar applications of machine learning in preventive healthcare, contributing to a reduction in the environmental burden associated with reactive and resource-intensive practices.

While the direct environmental impact is minimal, efforts should be made to optimize computational resources for energy efficiency and leverage sustainable energy sources whenever possible during model training and deployment.

5.3 Ethical Aspects

The development and deployment of machine learning models for stress and anxiety detection raises several ethical considerations that must be carefully addressed. Robust measures are necessary to protect the privacy and confidentiality of individuals' survey responses, adhering to data protection regulations. Rigorous testing and monitoring should ensure the models do not perpetuate biases or discriminate against certain groups. Efforts should be made to develop interpretable models or provide clear explanations, promoting transparency and trust. The models should complement, not replace, human expertise and clinical judgment, maintaining individual autonomy over healthcare decisions. Clear guidelines for ethical deployment, proper consent procedures, and educational resources are crucial. Mechanisms for accountability, oversight, and addressing potential issues should be established. By proactively addressing these ethical aspects, the benefits of machine learning can be realized while mitigating risks and upholding ethical principles.

5.4 Sustainability Plan

To ensure long-term sustainability and continuous improvement, a comprehensive plan encompassing data management, model maintenance, and resource allocation is essential. Continuous data collection, curation, and integration of new survey data should be systematic while adhering to privacy and quality standards. Regular monitoring, evaluation, and updating of models should maintain accuracy and relevance. Computational resources should be optimized for energy efficiency, leveraging renewable sources or exploring energy-efficient hardware. Collaborations with stakeholders can facilitate knowledge sharing, best practices, and potential enhancements. Securing sustainable funding sources and allocating appropriate resources is crucial. Staying abreast of evolving regulations and ethical guidelines, ensuring compliance with relevant standards is vital. By implementing a comprehensive plan addressing data, model maintenance, resource allocation, collaboration, funding, and ethical compliance, these machine learning models can continue evolving while adhering to responsible and sustainable practices.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future

6.1 Summary of the Study

The primary objective of this study was to develop an effective machine learning model for detecting stress and anxiety levels using survey data. The survey data consisted of responses to a series of questions designed to assess various behavioral and psychological indicators related to stress and anxiety, such as carelessness, lack of attention to detail, poor organizational skills, inability to focus or prioritize, forgetfulness, restlessness, edginess, mood swings, irritability, and risk-taking behaviors. A total of fourteen different machine learning models were evaluated, including Decision Tree, Random Forest, Logistic Regression, XGBoost, Support Vector Machine (SVM), CatBoost, Naive Bayes, Gradient Boosting, AdaBoost, Extra Trees, Bagging Classifier, Stochastic Gradient Descent, Passive Aggressive Classifier, and Ridge Classifier.

6.2 Conclusions

Upon analyzing the results, it was evident that certain models outperformed others in terms of accuracy and F1 score for stress and anxiety detection. The top-performing models were XGBoost, CatBoost, SVM, and Random Forest, with accuracy scores ranging from 0.915361 to 0.902821 and F1 scores ranging from 0.881057 to 0.865801.

The success of these models can be attributed to their ability to effectively handle complex and non-linear relationships present in the survey data, as well as their robustness to outliers and noise. XGBoost, in particular, is a powerful ensemble method that combines multiple weak decision tree models, allowing it to capture intricate patterns and interactions among the features. Similarly, Random Forest, another ensemble technique, leverages multiple decision trees and aggregates their predictions, often resulting in improved generalization and reduced overfitting.

CatBoost and SVM, although based on different underlying principles, also demonstrated excellent performance. CatBoost is a gradient boosting algorithm that efficiently handles categorical features and automatically incorporates ordered target encoding, which could have been beneficial for the survey data. SVM, a discriminative classifier, excels at finding optimal decision boundaries, making it well-suited for binary classification tasks like stress and anxiety detection.

This study demonstrated the potential of machine learning techniques to accurately detect stress and anxiety levels based on survey responses, which could have significant implications for mental health assessment and intervention. By identifying individuals at risk or experiencing high levels of stress and anxiety, appropriate support and resources can be provided promptly, potentially mitigating the negative impacts on overall well-being and productivity.

6.3 Implications for Further Study

While the current study has yielded promising results, there are several avenues for further exploration and improvement. Future research could investigate the incorporation of additional features or data sources, such as physiological measurements or behavioral data, to enhance the predictive power of the models. Additionally, exploring other advanced machine learning algorithms or deep learning techniques may uncover new insights and improve the accuracy of stress and anxiety detection. Furthermore, validating the generalizability of the models across different populations and domains would be crucial for broader applicability.

The findings of this study have practical implications for stress and anxiety management, mental health awareness, and early intervention strategies. By accurately detecting stress and anxiety levels, individuals can be identified and provided with appropriate support and resources before their condition escalates. Moreover, the models developed in this study could be integrated into mental health screening tools or personalized treatment plans, enabling more targeted and effective interventions.

Potential research directions could involve investigating the use of these models in real-world settings, such as workplaces, educational institutions, or healthcare facilities, to evaluate their performance and impact on mental well-being. Additionally, longitudinal studies could explore the utility of these models in monitoring stress and anxiety levels over time, enabling early detection of potential issues and facilitating timely interventions.

References:

[1] Robinson, Lisa. "Stress and anxiety." Nursing Clinics of North America 25, no. 4 (1990): 935-943.

[2] Daviu, Nuria, Michael R. Bruchas, Bita Moghaddam, Carmen Sandi, and Anna Beyeler. "Neurobiological links between stress and anxiety." Neurobiology of stress 11 (2019): 100191.

[3] Grillon, Christian, Roman Duncko, Matthew F. Covington, Lori Kopperman, and Mitchel A. Kling. "Acute stress potentiates anxiety in humans." Biological psychiatry 62, no. 10 (2007): 1183-1186.

[4] Shin, Lisa M., and Israel Liberzon. "The neurocircuitry of fear, stress, and anxiety disorders." Focus 9, no. 3 (2011): 311-334.

[5] Medvedev, Alexei, Natalia Igosheva, Michele Crumeyrolle-Arias, and Vivette Glover. "Isatin: role in stress and anxiety." Stress 8, no. 3 (2005): 175-183.

[6] Mogg, Karin, Andrew Mathews, Carol Bird, and Rosanne Macgregor-Morris. "Effects of stress and anxiety on the processing of threat stimuli." Journal of personality and social psychology 59, no. 6 (1990): 1230.

[7] Putwain, David. "Researching academic stress and anxiety in students: some methodological considerations." British educational research journal 33, no. 2 (2007): 207-219.

[8] Rada, Robert E., and Charmaine Johnson-Leong. "Stress, burnout, anxiety and depression among dentists." The Journal of the American Dental Association 135, no. 6 (2004): 788-794.

[9] Bouayed, Jaouad, Hassan Rammal, and Rachid Soulimani. "Oxidative stress and anxiety: relationship and cellular pathways." Oxidative medicine and cellular longevity 2, no. 2 (2009): 63.

[10] Beehr, Terry A., and Joseph E. McGrath. "Social support, occupational stress and anxiety." Anxiety, stress, and coping 5, no. 1 (1992): 7-19.

[11] Haller, J., and N. Bakos. "Stress-induced social avoidance: a new model of stress-induced anxiety?." Physiology & behavior 77, no. 2-3 (2002): 327-332.

[12] Maes, Michael, Cai Song, Aihua Lin, Raf De Jongh, An Van Gastel, Gunter Kenis, Eugene Bosmans et al. "The effects of psychological stress on humans: increased production of pro-inflammatory cytokines and Th1-like response in stress-induced anxiety." Cytokine 10, no. 4 (1998): 313-318.

[13] Kikusui, Takefumi, James T. Winslow, and Yuji Mori. "Social buffering: relief from stress and anxiety." Philosophical Transactions of the Royal Society B: Biological Sciences 361, no. 1476 (2006): 2215-2228.

[14] Bremner, J. Douglas, John H. Krystal, Steven M. Southwick, and Dennis S. Charney. "Noradrenergic mechanisms in stress and anxiety: I. Preclinical studies." Synapse 23, no. 1 (1996): 28-38.

[15] Hovatta, Iiris, Juuso Juhila, and Jonas Donner. "Oxidative stress in anxiety and comorbid disorders." Neuroscience research 68, no. 4 (2010): 261-275.

[16] Kumari, Archana, and Jagrati Jain. "Examination stress and anxiety: A study of college students." Global Journal of Multidisciplinary Studies 4, no. 1 (2014): 31-40.

[17] Vettore, M. V., A. T. T. Leão, A. M. Monteiro Da Silva, R. S. Quintanilha, and G. A. Lamarca. "The relationship of stress and anxiety with chronic periodontitis." Journal of clinical periodontology 30, no. 5 (2003): 394-402.

[18] Fedoce, Alessandra das Gracas, Frederico Ferreira, Robert G. Bota, Vicent Bonet-Costa, Patrick Y. Sun, and Kelvin JA Davies. "The role of oxidative stress in anxiety disorder: cause or consequence?." Free radical research 52, no. 7 (2018): 737-750.

[19] Cherry, Nicola. "Stress, anxiety and work: A longitudinal study." Journal of occupational psychology 51, no. 3 (1978): 259-270.
[20] Kurebayashi, Leonice Fumiko Sato, Juliana Miyuki do Prado, and Maria Julia Paes da Silva. "Correlations between stress and anxiety levels in nursing students." Journal of Nursing Education and Practice 2, no. 3 (2012): 128-134.
[21] Seo, Dongju, Aneesha Ahluwalia, Marc N. Potenza, and Rajita Sinha. "Gender differences in neural correlates of stress-induced anxiety." Journal of neuroscience research 95, no. 1-2 (2017): 115-125.
[22] Viveros, M. P., Eva M. Marco, and Sandra E. File. "Endocannabinoid system and stress and anxiety responses." Pharmacology Biochemistry and Behavior 81, no. 2 (2005): 331-342.
[23] Johnson, Philip L., Andrei Molosh, Stephanie D. Fitz, William A. Truitt, and Anantha Shekhar. "Orexin, stress, and anxiety/panic states." Progress in brain research 198 (2012): 133-161.
[24] Giannakakis, Giorgos, Matthew Pediaditis, Dimitris Manousos, Eleni Kazantzaki, Franco Chiarugi, Panagiotis G. Simos, Kostas Marias, and Manolis Tsiknakis. "Stress and anxiety detection using facial cues from videos." Biomedical Signal Processing and Control 31 (2017): 89-101.
[25] Priya, Anu, Shruti Garg, and Neha Prerna Tigga. "Predicting anxiety, depression and stress in modern life using machine learning algorithms." Procedia Computer Science 167 (2020): 1258-1267.
[26] Kumar, Prince, Shruti Garg, and Ashwani Garg. "Assessment of anxiety, depression and stress using machine learning models." Procedia Computer Science 171 (2020): 1989-1998.
[27] Bhatnagar, Shaurya, Jyoti Agarwal, and Ojasvi Rajeev Sharma. "Detection and classification of anxiety in university students through the application of machine learning." Procedia Computer Science 218 (2023): 1542-1550.
[28] Fatima, Asra, Ying Li, Thomas Trenholm Hills, and Massimo Stella. "Dasentimental: Detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning." Big Data and Cognitive Computing 5, no. 4 (2021): 77.
[29] Bobade, Pramod, and M. Vani. "Stress detection with machine learning and deep learning using multimodal physiological data." In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 51-57. IEEE, 2020.
[30] Tasnim, Mashrura, Ramon Diaz Ramos, Eleni Stroulia, and Luis A. Trejo. "A Machine-Learning Model for Detecting Depression, Anxiety, and Stress from Speech." In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7085-7089. IEEE, 2024.
[31] Richter, Thalia, Barak Fishbain, Andrey Markus, Gal Richter-Levin, and Hadas Okon-Singer. "Using machine learning-based analysis for behavioral differentiation between anxiety and depression." Scientific reports 10, no. 1 (2020): 16381.
[32] Sergio, Wagno Leão, Victor Ströele, Mário Dantas, Regina Braga, and Douglas Dyllon Macedo. "Enhancing well-being in modern education: A comprehensive eHealth proposal for managing stress and anxiety based on machine learning." Internet of Things 25 (2024): 101055.
[33] Singh, Srishti, Harsh Gupta, Prashant Singh, and Arun Prakash Agrawal. "Comparative Analysis of Machine Learning Models to Predict Depression, Anxiety and Stress." In 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), pp. 1199-1203. IEEE, 2022.
[34] Khullar, Vikas, Raj Gaurang Tiwari, Ambuj Kumar Agarwal, and Soumi Dutta. "Physiological signals based anxiety detection using ensemble machine learning." In Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021, pp. 597-608. Springer Singapore, 2022.

[35] Reddy, U. Srinivasulu, Aditya Vivek Thota, and A. Dharun. "Machine learning techniques for stress prediction in working employees." In 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-4. IEEE, 2018.

[36] Katiyar, Kalpana, Hera Fatma, and Simran Singh. "Predicting Anxiety, Depression and Stress in Women Using Machine Learning Algorithms." In Combating Women's Health Issues with Machine Learning, pp. 22-40. CRC Press, 2024.

[37] El Morr, Christo, Manar Jammal, Imad Bou-Hamad, Sahar Hijazi, Dinah Ayna, Maya Romani, and Reem Hoteit. "Predictive Machine Learning Models for Assessing Lebanese University Students' Depression, Anxiety, and Stress During COVID-19." Journal of Primary Care & Community Health 15 (2024): 21501319241235588.

[38] Trivedi, Naresh Kumar, Raj Gaurang Tiwari, Deden Witarsyah, Vinay Gautam, Alok Misra, and Ryan Adhitya Nugraha. "Machine Learning Based Evaluations of Stress, Depression, and Anxiety." In 2022 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS), pp. 1-5. IEEE, 2022.

[39] Malik, Shahid Shabeer, and Aneeque Khan. "Anxiety, Depression and Stress prediction among College Students using Machine Learning Algorithms." In 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), pp. 1-5. IEEE, 2023.

[40] Daadaa, Yassine. "EmotionNet: Dissecting Stress and Anxiety Through EEG-based Deep Learning Approaches." International Journal of Advanced Computer Science & Applications 15, no. 1 (2024).

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title: STRESS AND ANXIETY EARLY DETECTION WITH MACHINE LEARNING APPROACH USING SURVEY DATA.

Student ID: 201-15-14196

CO Description For FYDF

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to Detecting early stress and anxiety for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish these goals for the FYDP	**PO4**
CO4	Develop and validate a comprehensive machine learning-based system for the objective and data-driven detection of stress and anxiety, ensuring data integrity and ethical considerations are maintained. Utilize diverse machine learning models to generalize individual-specific patterns and ensure robust data handling and privacy measures throughout the project lifecycle.	**PO11**
Phase -II		
CO5	Create and develop systems or processes that meet specific needs, ensuring they follow public health and safety standards and consider cultural, socioeconomic, and environmental factors.	**PO3**
CO6	Select and use the right methods, tools, and modern technologies to solve complex problems, including prediction and modeling, while respecting relevant limitations.	**PO5**
CO7	Consider societal, health, safety, legal, and cultural factors, and understand your responsibilities in professional practice to solve problems logically.	**PO6**
CO8	Evaluate the long-term sustainability and impact of engineering projects on society and the environment	**PO7**
CO9	Follow ethical principles and professional standards throughout the project.	**PO8**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (With Knowledge Profile)	References
	EP1: Depth of Knowledge required		CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project demonstrates fundamental engineering principles (K3) by utilizing machine learning techniques, including Support Vector Machines (SVM), XGBoost, CatBoost, and other algorithms, to analyze survey response data for detecting stress and anxiety levels. The project showcases specialist knowledge (K4) by developing an objective, data-driven approach, training and evaluating 13 different machine learning models to identify individuals at risk. This automated system aims to provide timely intervention and treatment, significantly improving mental health outcomes and overall well-being.	**Page no:** [9] **Section:** [1.1]
				The project applies engineering practice and design (K5) through a comprehensive data collection procedure using both online and offline methods to ensure a diverse and representative sample. The project addresses engineering practice and technology (K6) by employing a diverse set of machine learning algorithms, including Support Vector Machines (SVM), XGBoost, and CatBoost, to analyze survey response data for detecting stress and anxiety levels.	**Page no:** [27] **Section:** [3.2]
				This project ensures K8 (Research Literature) by synthesizing insights from recent studies on stress and	**Page no:** **[23-30]**

				anxiety detection, employing a comprehensive set of machine learning algorithms to analyze survey data, and demonstrating a deep understanding of current methodologies and their applications.	**Section: [2.2, 2.3]**
2.		**Yes**		This project addresses EP-2 by recognizing the hurdles in detecting stress and anxiety, including limitations of traditional methods and the complexities of integrating machine learning techniques. Through comparative analysis, it confronts challenges in data quality, individual variability, and ethical considerations, offering insights for refining methodologies and improving detection accuracy.	**Page no: [30-31] Section: [2.4, 2.5]**
3.		**Yes**		This project addresses EP-3 by meticulously comparing experimental outcomes, highlighting machine learning as the chosen significant solution for enhancing stress and anxiety detection approaches.	**Page no: [37-48] Section: [4.2]**
4.		**Yes**		This project's interdisciplinary approach extends beyond computer science and engineering, impacting stress and anxiety detection methods EP-4.	**Page no: [23-31] Section: [2.2, 2.4]**
5.		**No**		N/A	**N/A**
6.		**No**		N/A	**N/A**
7.		**Yes**		This project's comprehensive approach addresses high-level problems by integrating various components across data collection, statistical analysis, and proposed methodology, ensuring a holistic solution to complex challenges in stress and anxiety detection, which	**Page no: [26-34] Section: [3.2, 3.3,**

				ensures EP-7.	**3.4]**

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, various machine learning frameworks, annotated datasets, and ethical considerations to ensure systematic research and contribute to advancements in stress and anxiety detection	**Page no:** [33-35] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A

4.	**EA4: Consequences for society and the environment**	Yes	This project contributes to society by improving stress and anxiety detection methods, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for data privacy.	**Page no:** [50-51] **Section:** [5.1, 5.2, 5.4]
5.	**EA-5: Familiarity**	**Yes**	This project expands upon existing research by examining a novel approach in detecting stress and anxiety through machine learning models demonstrated through preliminary methodologies and a comprehensive comparative analysis, offering new insights into the field.	**Page no:** [23-29] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses CO4 by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal project execution throughout the research lifecycle.	**Page no:** [20] **Section:** [1.6]
2	CO9	Yes	The project demonstrates adherence to ethical principles by prioritizing privacy, obtaining informed consent, and transparently documenting the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced classification technologies, which comply with CO9.	**Page no:** [51] **Section:** [5.3]
3	CO10	No	N/A	N/A

4	CO12	Yes	The project's dedication to continuous learning (CO12) and adaptation within the dynamic technological landscape is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges..	**Page no:** [32-36, 37-48] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]

Bandhan

ORIGINALITY REPORT

19%	15%	7%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	fastercapital.com Internet Source	2%
2	Submitted to Napier University Student Paper	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	www.braze.com Internet Source	1%
5	www.coursehero.com Internet Source	1%
6	deepblue.lib.umich.edu Internet Source	1%
7	Submitted to Liverpool John Moores University Student Paper	1%
8	www.researchgate.net Internet Source	<1%
9	www.utupub.fi Internet Source	<1%