ANALYZING IMPACT OF PROLONGED USE OF SOCIAL MEDIA AMONG UNIVERSITY STUDENTS

By

Md. Nurnabi
210221014
A.S.M Walee Ullah Naym
210221002
Md. Mojibar Rahman
210221007
Alamgir Hossen
200321001
Trijita Khisa
210221012

A thesis submitted to the Department of Computer Science & Engineering in partial fulfillment of the requirements for the degree of
BSc. in Engineering.

Department of Computer Science & Engineering
European University of Bangladesh
May 2025

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at European University of Bangladesh.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student’s Full Name & Signature:

Md. Nurnabi 210221014	**A.S.M Walee Ullah Naym** 210221002
Md. Mojibar Rahman 210221007	**Alamgir Hossen** 200321001
Trijita Khisa 210221012	

Approval

The thesis titled "Analyzing impact of prolonged use of social media among university students" Submitted by

1. Md. Nurnabi (210221014)
2. A.S.M Walee Ullah Naym (210221002)
3. Md. Mojibar Rahman (210221007)
4. Alamgir Hossen (200321001)
5. Trijita Khisa (210221012)

of 12^{th} semester, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of BSc. In Engineering on 31-05-2025.

Examining Committee:

Supervisor:

____
Khandaker Iftakher Ahmed
Lecturer, Department of CSE,
EUB.

Project/Thesis Coordinator:

____
Fariha Nusrat
Lecturer, Department of CSE,
EUB.

Departmental Head:
(Chair)

____
Chowdhury Sajadul Islam
Associate Professor and Chairman,
Department of CSE,
EUB.

Ethics Statement

This research was conducted in accordance with standard academic ethical guidelines to ensure the rights, privacy, and dignity of all participants were respected throughout the study. Informed consent was obtained from all participants prior to data collection. They were clearly informed about the purpose of the research, their right to withdraw at any time, and how their responses would be used.

All data collected through surveys, interviews, and focus group discussions were kept strictly confidential and anonymized to protect participants' identities. No personal or sensitive information was disclosed or shared with any third parties. The study did not involve any physical or psychological harm, and care was taken to avoid sensitive or intrusive questions.

Ethical approval, if required, was sought from the appropriate academic or institutional review body. This research was undertaken with the intention to contribute positively to academic knowledge and student well-being, without any form of coercion, bias, or exploitation.

Abstract/ Executive Summary

The pervasive integration of social media into the daily lives of university students of Bangladesh has raised significant concerns about addiction and its repercussions on academic, psychological, and social well-being. This study investigates the prevalence, drivers, and consequences of prolonged social media use among students in Bangladesh, leveraging a mixed-method approach combining quantitative surveys (N=899) and qualitative interviews. Against a backdrop of rapid digitalization fueled by affordable smartphones, widespread internet access, and unregulated platform engagement the research identifies addiction patterns through behavioral metrics (e.g., compulsive checking, withdrawal symptoms) and machine learning models (Random Forest, Naïve Bayes, Logistic Regression).

Key findings reveal that 65% of participants exhibit addiction-like behaviors, with 59% reporting >5 hours of daily use, primarily on Facebook, YouTube, TikTok and Instagram. Academic disruptions are stark: 42% of students attribute declining grades to procrastination, while 38% experience sleep deprivation due to late-night browsing. Mental health challenges, including anxiety (68%) and social comparison tendencies (55%), are strongly correlated with excessive use. Gender and faculty disparities emerge, with male students (59%) and Engineering majors (39%) dominating high-usage cohorts. Machine learning analysis underscores Logistic Regression as the most effective predictor (39% accuracy), outperforming Naïve Bayes (32%) and Random Forest (23%), while Linear Regression proved invalid ($R^2 = -0.01$).

Cultural factors, such as familial pressures and peer validation, exacerbate dependency, compounded by institutional gaps in digital literacy education. The study advocates for policy reforms, including campus-based mental health support, screen-time awareness campaigns, and curriculum-integrated digital wellness programs. By bridging empirical gaps in the Global South context, this research underscores the urgency of balancing digital engagement with academic and psychological resilience, offering actionable strategies for stakeholders in Bangladeshi higher education.

Keywords: “Social media addiction”, “University students”, “Bangladesh”, “Academic performance”, “Mental health”, “Digital behavior”, “Smartphone use”, “Digital literacy” and “Online dependency”.

Dedication

We dedicate this work to our parents, whose unwavering support and encouragement have been our greatest source of strength. Their constant belief in us has pushed us to pursue our studies with determination and resilience. Through every challenge, they have stood by our side, inspiring us to work hard and never lose sight of our goals. This achievement is as much ours as it is theirs. Our study in a successful manner and for their unconditional support they did for us to achieve this level of academic endeavor and the support and guidance throughout our lives. Our Allah rests their souls in the heaven and be merciful to them as they were to us in our childhood (Amen).

Acknowledgement

At first, all praises and thanks are due to our creator Allah (SWT), the creator, sustainer and the most merciful, who gave us the courage, endurance, willingness and ability to complete this paper successfully.

This report forms the thesis of our graduation research paper that is conducted in partial fulfillment of the requirements for the bachelor degree, department of Computer Science and Engineering. We have learned a lot from this study and it has encouraged us to cheer up and work harder in the future. We are also gratitude to all those who have directly or indirectly contributed in different manners to this thesis paper and we would like to thank and acknowledge a number of people. Secondly, we would like to thank our beloved parents as well as our teacher's. Because they made us, with so much love, care and discipline. We thank them for all they have done for us so far and their unconditional support through our study and personal career.

Secondly, we would like to thank to all members of our families who also supported us. Thirdly, we would like to express our great gratitude to our supervisor, Khandaker Iftakher Ahmed for his abundant help, invaluable assistance, support and guidance.

Finally, we would like to thank our CSE department, the European University of Bangladesh, for providing us with such a beautiful environment to conduct research. We have benefited greatly from having adequate lab facilities to conduct research in our department, so I express my gratitude to the department authorities.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AP	Academic Performance
DA	Dating Apps
DC	Demographic Characteristics
DD	Data-Driven (approach/methods)
EO	Emotional Outcomes
FB	Facebook
GT	Gratification Theory
IG	Instagram
SA	Statistical Analysis
EDA	Exploratory Data Analysis
LR	Logistic Regression
MH	Mental Health
ML	Machine Learning
NB	Naïve Bayes
POL	Policy (interventions/guidelines)
RF	Random Forest
DC	Data collection
DCM	Data collection methods
SD	Sleep Disturbances
SMA	Social Media Addiction
SMU	Social Media Usage
TM	Time Management
TT	TikTok
US	University Students
WA	WhatsApp
YT	YouTube
SSRE	Selective Serotonin Reuptake Enhancer
SSTOT	Total Sum of Squares

Glossary

Addiction:	A psychological or behavioral dependence on a substance or activity, often leading to negative consequences and difficulty in controlling use. In this study, it refers to compulsive social media use.
Digital behavior:	The way individuals interact with digital devices and platforms, including usage patterns, habits, and emotional responses.
Facebook/Instagram/ TikTok /YouTube:	Popular social media platforms widely used by university students for communication, entertainment, and self-expression.
Mental health:	A person's emotional, psychological, and social well-being, which can be affected by excessive social media use.
Mixed-method research:	A research approach that combines both quantitative (numerical data) and qualitative (descriptive data) methods to gain a fuller understanding of a topic.
Screen time:	The amount of time spent using devices such as smartphones, tablets, or computers, particularly for non-academic purposes.
Social comparison:	The act of comparing oneself to others, often seen on social media, which can affect self-esteem and well-being.
Social media:	Online platforms and applications that enable users to create, share, and interact with content and other users in real time.
University students:	Individuals enrolled in undergraduate or postgraduate programs at public or private universities in Bangladesh.

.

Chapter 1

Introduction

Social media has revolutionized the way people communicate, share information, and interact with the world. Over the past decade, platforms such as Facebook, Instagram, TikTok, YouTube and twitter have become an integral part of everyday life, especially among youth. University students, in particular, are among the most active social media users, often engaging with these platforms for both academic and non-academic purposes. While social media offers various benefits, including staying connected with friends and accessing educational content, excessive and uncontrolled use has given rise to a new form of behavioral concern social media addiction.

In Bangladesh, the growth of internet accessibility, smartphone ownership, and affordable data packages has led to a surge in social media usage, particularly among university students. This demographic, being in a transitional phase of life marked by emotional, social, and academic challenges, is highly vulnerable to addictive patterns of online behavior. Studies have shown that when social media use becomes compulsive or interferes with daily functioning, it can negatively affect students' academic performance, mental health, sleep quality, and real-life relationships.

Social media addiction is not yet officially recognized as a clinical disorder by diagnostic manuals such as the dsm-5, but it shares many symptoms with other behavioral addictions, such as mood modification, salience, tolerance, withdrawal, conflict, and relapse. In the context of Bangladesh, where discussions on digital well-being are still emerging, there is limited empirical research focused specifically on the impact of social media overuse on university students [1].

This study seeks to fill that gap by exploring the extent, nature, and effects of social media addiction among university students in Bangladesh. Through a mixed-method research design, the study aims to identify common behavioral patterns, contributing psychological and social factors, and the broader academic and health-related consequences. The ultimate goal is to raise awareness, inform policy, and propose practical strategies for prevention and intervention within Bangladeshi higher education institutions [2].

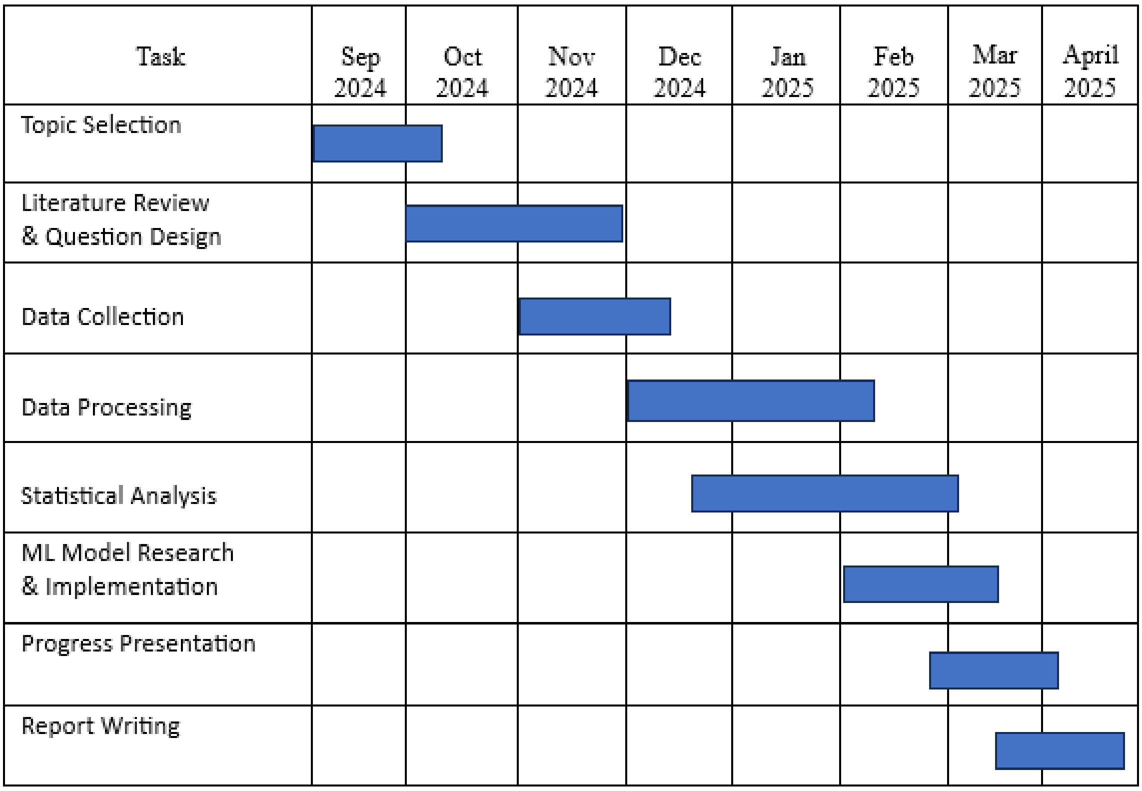

Figure 1: Timeline of the Research

1.1 Motivation

In recent years, social media has become an essential part of student life in Bangladesh. While platforms like Facebook, Instagram, and TikTok offer communication, entertainment, and learning opportunities, their excessive use has led to concerning behavioral patterns. Many university students spend several hours daily on these platforms, often at the cost of academic focus, mental well-being, and real-life interactions. This trend is growing rapidly, yet it remains under-researched in the Bangladeshi context.

As digital access continues to expand with affordable smartphones and internet services, students are increasingly vulnerable to social media addiction. However, most universities lack awareness programs, support systems, or guidance on healthy digital use. This gap in understanding and intervention motivated this research to explore how deeply social media addiction affects students in Bangladesh and to provide practical recommendations for early awareness, prevention, and support [3].

1.2 Objectives

The primary objective of this study is to explore the extent, causes, and consequences of social media addiction among university students in Bangladesh. The research aims to:

- To find out how much time university students in Bangladesh spend on social media they are.
- To acknowledge what social and mental factors cause students to become prolonged use social media.
- To identify which faculty students, use social media more.

1.3 Problem Statement

In recent years, social media has become deeply integrated into the daily lives of university students in Bangladesh. While it offers many benefits, such as communication, learning, and entertainment, excessive use has led to growing concerns about social media addiction. Many students spend significant amounts of time online, which can negatively affect their academic performance, mental health, and social relationships.

Despite the increasing awareness of these issues, there is limited research in the Bangladeshi context that explores the extent of this addiction, it’s root causes, and practical ways to manage or reduce it. Without proper understanding and intervention, social media addiction may continue to impact students' wellbeing and future potential. This study aims to analyze the depth of this problem, identify contributing factors, and evaluate strategies that can help students use social media more responsibly [4].

1.4 Methodology

The Research Flowchart outlines the structured process for analyzing the impact of prolonged social media use among university students. It starts with defining the research topic, followed by data collection using Google Forms and exporting responses to Excel. The data preprocessing phase ensures the dataset is clean and formatted correctly. Next, statistical analysis is conducted using Exploratory Data Analysis (EDA) and basic statistical methods.

Following this, machine learning models Random Forest, Linear Regression, Logistic Regression, and Naive Bayes are trained and tested to derive insights. Visualization techniques such as bar charts, line graphs, and pie charts help present findings effectively. The process concludes with report writing, summarizing methodology, results, and implications, leading to the final conclusion that provides key takeaways.

- Conduct surveys among university students to measure social media usage time and dependency levels.

- Host focus group discussions to delve into personal experiences and coping mechanisms.

- Analyze individual cases of students who have successfully overcome social media addiction.

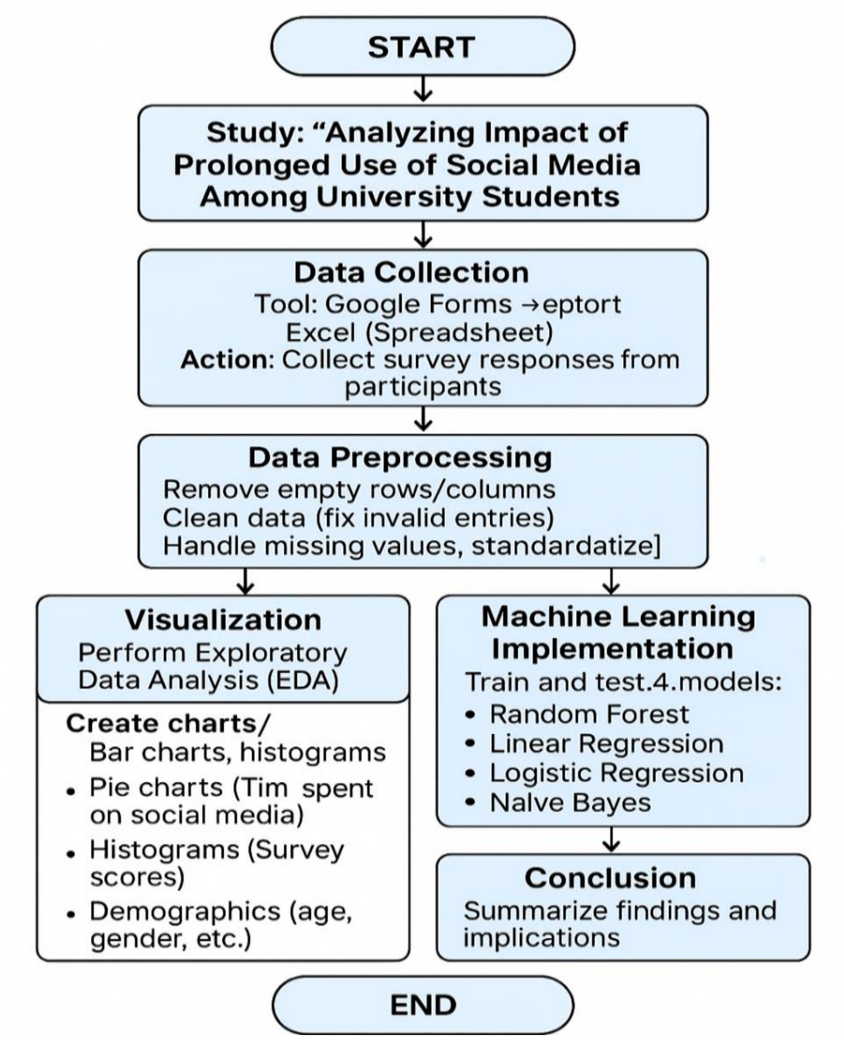

Figure: 1.4.1 Methodology

1.5 Scope of Study

This study focuses on understanding the nature and impact of social media addiction specifically among university students in Bangladesh. The research is limited to a selected number of public and private universities located in major cities such as Dhaka, Chittagong, Rajshahi and Sylhet. It targets undergraduate and postgraduate students who are active users of popular social media platforms like Facebook, Instagram, TikTok, and Snapchat.

The study examines patterns of social media usage, factors contributing to addictive behavior, and the effects on students' academic performance, mental health, and social life. Both quantitative (survey-based) and qualitative (interview/focus group) data collection methods are used to ensure a comprehensive understanding.

However, the study does not cover primary or secondary school students, non-student populations, or individuals outside Bangladesh. It also does not address internet addiction in general or addiction related to online gaming or other digital content.

1.6 Background Study

In recent years, social media has become an essential part of student life, particularly among university students in Bangladesh. Platforms like Facebook, Instagram, TikTok, YouTube and various messaging or dating apps are widely used not only for entertainment and communication but also for academic collaboration and self-expression. The increasing availability of affordable smartphones and internet packages has further accelerated digital engagement among students across the country.

Students often spend several hours daily on social media, with research indicating average usage ranging from 3 to 5 hours per day. While this trend reflects a growing digital culture, it also raises concerns about the effects of prolonged use on students' daily lives. According to the uses and gratification theory, individuals turn to social media to fulfill specific emotional and psychological needs such as relieving stress, seeking social approval, and staying connected. Although these platforms offer clear benefits, overreliance on them can have adverse consequences.

Many students experience challenges such as poor time management, sleep disturbances, reduced focus, and increased anxiety as a result of excessive screen time. In academic settings, this can translate into decreased performance, lack of motivation, and difficulty maintaining a balanced routine. Students in major cities like Dhaka, Chittagong, Raj Shahi, and Sylhet are especially exposed due to easy access to technology and dense digital environments. Despite the growing relevance of this issue, there remains a gap in research that specifically examines the patterns, causes, and impacts of prolonged social media use among university students in Bangladesh. This study aims to address that gap by exploring how these behaviors affect students' academic and personal lives, and by providing insights that can support early awareness, prevention, and institutional policy development [5].

1.7 Contribution

The research makes significant contributions by addressing the underexplored issue of prolonged social media use among university students in Bangladesh, offering a nuanced understanding of its academic, psychological, and cultural implications. The context of Bangladesh, a mixed-method approach combining surveys of 899 students and qualitative interviews reveals that 65% of participants display addiction-like behaviors, with 59% spending over 5 hours daily on platforms like Facebook, YouTube and Instagram. This compulsive usage is directly tied to stark academic disruptions: 42% of students blame declining grades on procrastination, while 38% face sleep deprivation from late-night browsing. The study also uncovers profound mental health challenges, with 68% reporting anxiety and 55% struggling with social comparison, underscoring the emotional toll of unchecked digital engagement.

Ultimately, the work transcends theoretical analysis by proposing actionable solutions tailored to Bangladesh's socio-digital landscape, such as integrating digital wellness into curricula, launching campus mental health initiatives, and advocating for policy reforms to regulate platform engagement. By contextualizing global theories like the Uses and Gratification framework within Bangladesh's rapid yet unregulated digitalization, the study not only fills empirical gaps but also provides a blueprint for balancing technological advancement with academic and psychological resilience in developing economies.

1.8 Organization

This thesis, titled "Analyzing the Impact of Prolonged Use of Social Media Among University Students" is structured to provide a comprehensive examination of the effects of social media usage. The organization is designed to ensure clarity, logical progression, and coherence in presenting the research findings.

Chapter 1: Introduction

This chapter outlines the research problem, objectives, significance, and scope of the study. It introduces key concepts related to social media and its impact on university students.

Chapter 2: Literature Review

A critical analysis of previous studies related to social media usage, its psychological and academic effects, and existing theories. This chapter establishes the theoretical framework for the study.

Chapter 3: Research Models

This section describes the research design, discuss about machine learning models like Random Forest, naive Bayes, Linear Regression and Logistic Regression sample population, and analytical techniques.

Chapter 4: Implementation and Result Analysis

Presentation of data collected from surveys, interviews, or experiments. This chapter interprets results based on statistical and thematic analysis, linking findings to existing literature.

Chapter 6: Conclusion

A summary of key findings, conclusions drawn from the study, and recommendations for students, educators, and policymakers. It also suggests areas for future research. This structured approach ensures a logical flow, making it easier for readers to grasp the research insights.

Chapter 2

Literature Review

The rapid proliferation of social media platforms has dramatically reshaped communication, information exchange, and social interactions, particularly among university students. Several studies have explored the phenomenon of social media addiction, a behavioral dependency characterized by excessive use that interferes with daily life activities, academic performance, and psychological well-being.

Prevalence and associated factors of internet addiction among young adults in Bangladesh.

Internet addiction (IA), first introduced by Kimberly Young in 1996, refers to excessive or poorly controlled internet use that leads to distress or functional impairment. Although not officially classified in the DSM-5, IA is increasingly recognized as a serious issue, especially among young adults. Global studies have linked IA to negative outcomes like academic problems, social isolation, and poor family relationships. Risk factors include younger age, male gender, lack of family support, excessive time online especially on social media and habits like smoking. Despite growing global concern, research on IA in Bangladesh remains limited, highlighting the need for local studies like that of Hassan et al. (2020) [6].

Psychometric assessment of the Bangla version of the Bergen Social Media Addiction Scale.

The study evaluates the reliability and validity of the Bangla adaptation of the Bergen Social Media Addiction Scale (BSMAS) for assessing social media addiction among Bangladeshi university students. Using data from 577 participants, the study employed Classical Test Theory (CTT), Item Response Theory (IRT), and network analysis to confirm the scale's single-factor structure and strong internal consistency. The findings indicate that the Bangla BSMAS is a reliable tool, free from gender or usage-duration bias, and effective in identifying core symptoms of social media addiction, such as relapse and salience. While its applicability is limited to university students, the scale demonstrates significant potential for research and clinical interventions targeting social media dependency in young Bangladeshi adults [7].

A Study on Social Media Addiction of Young People during COVID-19 in Bangladesh.

The paper examines the impact of social media addiction on the mental health of young people in Bangladesh during the COVID-19 pandemic. Through online interviews with 520 university students, the study highlights a high prevalence of social media addiction, with 70% of participants exhibiting signs of addiction. This overuse correlated strongly with increased anxiety, depression, and other mental health challenges, with 80% reporting heightened stress and feelings of helplessness. The research identifies significant negative effects on physical and mental well-being and offers recommendations for mitigating addiction, including limiting usage, engaging in hobbies, and fostering direct interpersonal interactions [8].

Psychometric properties of three online-related addictive behavior instruments among Bangladeshi school-going adolescents.

The study assessed the psychometric properties of three tools—the Bangla versions of the Internet Gaming Disorder Scale-Short Form (IGDS9-SF), Gaming Disorder Test (GDT), and Bergen Social Media Addiction Scale (BSMAS)—for measuring online addictive behaviors among Bangladeshi school-going adolescents. Conducted through a cross-sectional survey with 428 participants, the study established the reliability, criterion validity, convergent validity, and construct validity of these instruments. The findings highlighted strong internal consistency and significant correlations between these tools and indicators like depression, anxiety, and internet usage, thus validating their utility in the Bangladeshi context for research and clinical applications [9].

Social Media, New Cultures, and New Threats: Impact on University Students in Bangladesh.

The study examines the dual impact of social media on university students in Bangladesh, highlighting its role in fostering new cultures while simultaneously introducing new threats. Conducted through surveys and case studies involving 217 students, the research reveals that social media has enabled effective communication, education, and entrepreneurship. However, it has also led to issues like dysfunctional relationships, cyberstalking, declining academic performance, and the proliferation of yellow journalism. The findings emphasize the need for greater awareness, stricter regulations, and ethical practices to mitigate the risks while maximizing the benefits of social media for personal and societal growth [10].

Association between flaunting behaviors on social media and among the general population in Bangladesh: A cross-sectional study.

The study explores the connection between social media flaunting behaviors and mental health in Bangladesh through a nationwide cross-sectional survey of 465 participants aged 18–60. It reveals a significant prevalence of loneliness (65.16%) and depressive symptoms (55.49%) among users, with factors like young age, male gender, being unmarried, and using social media for popularity contributing to these outcomes. Excessive social media use and self-expression trends, such as posting for validation or competing online, were linked to heightened risks of mental health issues. The findings emphasize the need for awareness, authentic online self-presentation, and policies to mitigate social media's psychological impact. The authors suggest further research with larger samples to better understand these dynamics [11].

Impact of social media usage on academic performance of university students: Mediating role of mental health under a cross-sectional study in Bangladesh.

The study investigates the impact of social media usage on the academic performance of university students in Bangladesh, focusing on the mediating role of mental health. Using data from 380 students and employing structural equation modeling, the research finds a statistically significant positive relationship between social media usage and both mental health and academic performance. It also identifies a mediating effect of mental health in the relationship between social media addiction and academic outcomes. While the findings highlight potential academic benefits of controlled social media use, they emphasize the importance of managing excessive use to mitigate mental health risks. The study contributes to understanding the nuanced effects of social media in educational contexts and offers implications for policymakers and university administrators [12].

Examining the benefits and drawbacks of social media usage on academic performance: a study among university students in Bangladesh.

This research investigates how social media affects academic outcomes for university students in Bangladesh, blending surveys (247 participants) and statistical methods. Results reveal social media aids learning through collaboration and resource access but harms performance when overused, causing distraction and reliance on technology. Popular platforms like Facebook and WhatsApp dominate usage. The research emphasizes the importance of policies that encourage balanced digital habits and informed internet use. However, reliance on self-reported data, a sample skewed toward younger students (99.6% under 24), and Bangladesh-specific contexts limit broader applicability. Insights remain valuable for shaping educational strategies in comparable settings [13].

Problematic Smartphone and Social Media Use Among Bangladeshi College and University Students Amid COVID-19: The Role of Psychological Well-Being and Pandemic Related Factors.

This study examines smartphone and social media addiction among 5,511 Bangladeshi college and university students during COVID-19 lockdowns. Findings reveal that excessive use correlates with anxiety, depression, poor sleep, and sedentary habits, particularly among younger, urban, and female students. Regression models identify psychological distress and inactivity as key drivers of dependency. While digital tools aided connectivity and information access, overuse risks addiction and academic disengagement. The study advocates for policies promoting digital literacy and balanced tech use. Limitations include reliance on self-reported data and a youth-centric sample, limiting broader applicability. Insights highlight urgent needs for mental health interventions in similar educational contexts post-pandemic [14].

Social media addiction & its effect on mental health among the private university students in Bangladesh.

This study investigates social media addiction and mental health impacts among private university students in Bangladesh via a questionnaire survey. Results indicate significant correlations between excessive social media use and mental health issues like anxiety and depression. Factors such as prolonged screen time and lack of offline engagement were linked to addiction. The study advocates for institutional policies promoting digital balance and mental health support. Limitations include reliance on self-reported data and a non-representative sample. Insights highlight the urgency of addressing tech overuse in academic settings to safeguard student well-being [15].

Chapter 3

Models

3.1 Random Forest

Random Forest is a powerful and widely used technique in data science that helps us make accurate predictions by analyzing complex datasets. At its core, this model is built using multiple decision trees, where each tree learns from a different subset of the data and makes its own independent prediction. Instead of relying on just one tree, Random Forest takes the average or weighted average of all the trees' predictions to produce a final result, making it more robust and reliable.

Because Random Forest is a supervised learning algorithm, it requires labeled data meaning the dataset contains known outcomes, which helps the model learn patterns effectively. The key advantage of this approach is that it creates multiple decision trees using randomly selected portions of the data.

Random Forest working principle

The process behind Random Forest regression is straightforward yet highly effective. It starts with the creation of multiple decision trees, each trained on different random subsets of the dataset. This technique, known as "bagging," helps prevent overfitting and ensures that the model captures diverse patterns within the data.

Once all the decision trees have been built, the model aggregates their predictions to produce a final output. By averaging these predictions, Random Forest improves reliability and minimizes biases, making it a great choice for scenarios where accuracy is crucial.

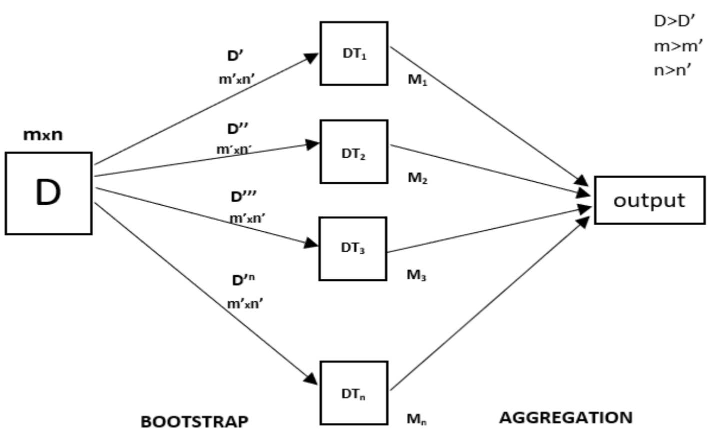

Figure 3.1.1: Working principle of Random Forest

Building a Random Forest Regression Model

Choose relevant variables to enhance model accuracy and predictive reliability. These features should be relevant to the outcomes, helping the model understand relationships and improve the accuracy of predictions. We also need to decide on the model's size, which will determine the number of trees in the forest and their depth. After these parameters are set, we train the model on our dataset using various techniques.

Evaluating Random Forest Regression Performance

Assess accuracy, analyze errors, and refine model for better predictions. It involves assessing its performance on unseen data points and measuring accuracy and precision. We can also use more advanced techniques, such as cross-validation, to ensure that our model is not overfitting or underfitting the data during training.

Interpreting Random Forest Regression Results

Interpret the model by analyzing feature importance and prediction accuracy. Determine important features and their relationships to enhance prediction accuracy. Use feature importance ranking and partial dependence plots to better understand the model's predictions and insights.

Time to use Random Forest Algorithm

Random Forest predicts continuous values like stock prices and market trends accurately. It efficiently handles large datasets and captures complex, non-linear relationships between variables. Averaging decision trees minimizes overfitting and enhances prediction accuracy. Unlike traditional linear models, it performs well with high-dimensional data, missing values, and categorical features, making it a reliable choice for businesses dealing with diverse datasets. Its ability to process varied data types while maintaining predictive strength makes it a valuable tool for financial forecasting, market analysis, and various other applications where precision is essential. Use Random Forest Regression for complex data with multiple inputs, ensuring accurate predictions while minimizing overfitting.

Random Forest is a powerful machine learning algorithm that builds multiple decision trees and combines their predictions for accuracy. It's widely used for classification and regression tasks, offering reliability by reducing overfitting and handling large datasets effectively.

Features of Random Forest

Some important features of this model are:

Speed: Random forests are faster and more efficient than many other machine learning techniques, like boosting. Since they require fewer parameters, they can process large datasets quickly while maintaining accuracy.

Accuracy: One of the biggest advantages of random forests is their ability to produce highly accurate predictions. They create multiple decision trees, each analyzing data points across different features, leading to well-classified results.

Feature Selection: Random forests help identify the most critical features in a dataset. By building numerous decision trees, the algorithm determines which variables contribute most to the predictive accuracy, making models more reliable.

Automation: This algorithm is easy to automate, making it a great choice for data scientists and analysts who need efficient and reliable results without manually adjusting multiple parameters.

Robustness: Random forests handle noisy or incomplete data exceptionally well. Even when datasets contain missing values or inconsistencies, the model still produces stable and precise predictions. This resilience makes it suitable for real-world applications.

Low Overfitting Risk: Unlike single decision trees, random forests reduce overfitting risks by generating multiple independent trees. Since no single tree dominates the outcome, the final prediction is more generalized and better suited for handling new data.

Additional Advantages

Diversity of Trees: Each decision tree within a random forest is unique. The variation between trees ensures diverse perspectives, improving overall model performance.

Handling High-Dimensional Data: Random forests are immune to the curse of dimensionality. Since they don’t require feature-specific assumptions, they efficiently process high-dimensional datasets without losing accuracy.

Parallel Processing: Each decision tree is built independently, allowing models to be constructed in parallel using CPUs, significantly speeding up computations.

3.2 Naïve Bayes

Naive Bayes Classifier: A simple powerful Algorithm

Naive Bayes classifiers are widely used for classification tasks in machine learning. They rely on Bayes' Theorem to calculate probabilities, making them highly efficient for applications such as spam detection, sentiment analysis, and text classification.

Key Features

A probabilistic approach that predicts based on likelihood rather than complex computations. Assumes features are independent, meaning each one contributes to classification without being influenced by others. Fast computation due to its simple structure and minimal parameters, allowing quick predictions.

Types of Naive Bayes Models

Gaussian Naive Bayes:

Ideal for continuous data, following a normal distribution that forms a smooth, bell-shaped curve. Often used in finance and medical diagnostics, where feature values follow Gaussian patterns.

Multinomial Naive Bayes:

Ideal for text classification, understanding word patterns and organizing documents efficiently. Commonly applied in spam filtering and document classification.

Bernoulli Naive Bayes:

Works with binary data, ideal for tasks like fraud detection or spam classification. Predicts based on presence or absence of features rather than their frequency.

Multinomial Naive Bayes

Multinomial Naive Bayes helps classify text by analyzing word frequencies, making it ideal for applications like spam detection and topic categorization. Bernoulli Naive Bayes, however, focuses on whether words are present or absent, rather than how often they appear. This makes it useful for sentiment analysis and fraud detection. Both models play a crucial role in organizing and understanding textual data, helping businesses and researchers make informed decisions based on document classification.

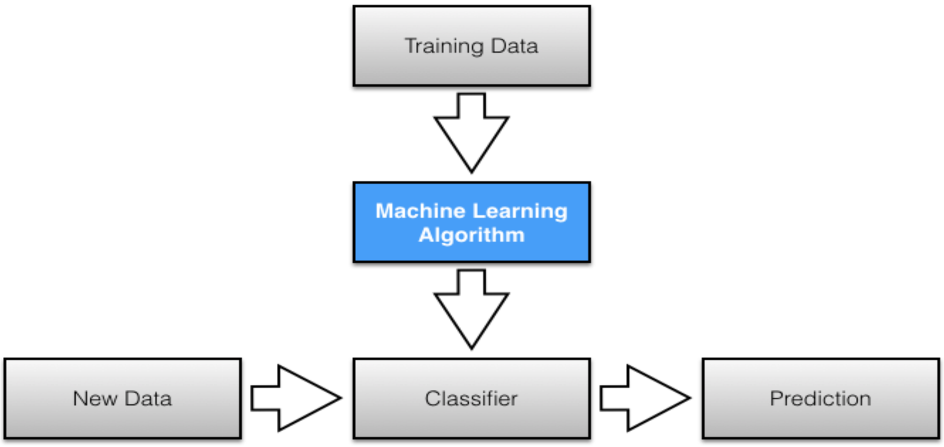

Figure 3.2.1: Naïve Bayes working Principle.

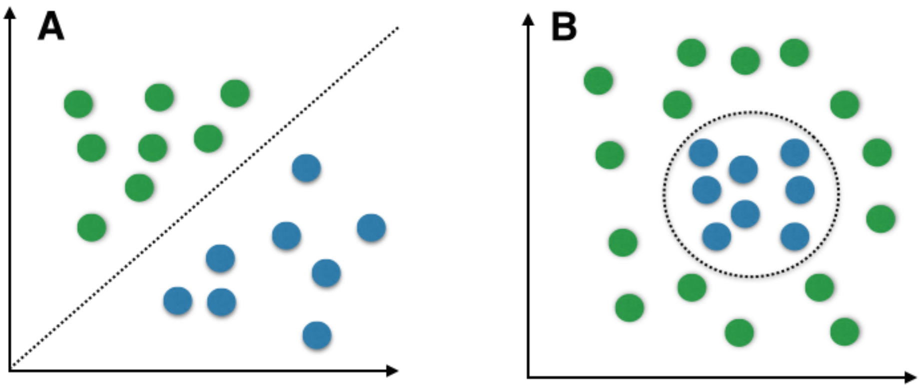

Figure 3.2.2: Linear vs. non-linear problems.

Posterior Probabilities

In order to understand how naive Bayes classifiers work, we have to briefly recapitulate the concept of Bayes' rule. The probability model that was formulated by Thomas Bayes (1701-1761) is quite simple yet powerful; it can be written down in simple words as follows:

$$\text{Porterior Probability } = \frac{\text{Conditional Probability. Prior Probability}}{\text{Evidence}}$$

Bayes' Theorem:

At its core, Bayes' Theorem is:

$$P(C\ |X) = \frac{\mathrm{P(X|C).\,P(C)}}{\mathrm{P(X)}}$$

Where:

- P(C|X): Posterior probability of class C given data X
- P(X|C): Likelihood of data X given class C
- P(C): Prior probability of class C
- P(X): Probability of data X

Naive Bayes Theorem in Classification:

For a classification task with features x_1, x_2, ..., x_n, Naive Bayes assumes:

$$\mathrm{P(C}\ |x1, x2, \ldots xn) \propto P(C).\prod_{i=1}^{n}.P(xi\ |C)$$

This means the model selects the class CCC that maximizes this probability — that is, the class that is most probable given the observed data.

Naive Bayes Performance:

Naive Bayes often performs surprisingly well for a wide range of tasks, despite its simplistic assumptions. Here's a breakdown of its performance characteristics:

Strengths

1. Speed & Efficiency
   - Very fast to train and predict, even on large datasets.
   - Low computational cost (no complex optimization or parameter tuning needed).
2. Works Well with High-Dimensional Data
   - Particularly effective in text classification (spam detection, sentiment analysis, document categorization).
3. Robust to Irrelevant Features
   - Since it treats features independently, irrelevant features have less impact.
4. Performs Well with Small Datasets
   - Can generalize well even when training data is limited, assuming the independence assumption holds reasonably well.

Weaknesses

1. Strong Independence Assumption:
   - Real-world data often includes correlations between features, which Naive Bayes ignores. This can reduce accuracy.
2. Zero-Frequency Problem:
   - If a category/word never appeared in the training set for a given class, it gets a probability of zero. This can be handled with Laplace smoothing.
3. Less Flexible for Complex Patterns:
   - May be outperformed by more complex models (e.g., Random Forest, SVM, Neural Networks) when feature interactions matter.

Assumptions of Naïve Bayes

Feature Independence: All features are assumed to be independent given the class label.

1. Equal Feature Importance: Each feature contributes equally and independently to the outcome.
2. Correct Distribution: Assumes feature values follow a known distribution (e.g., Gaussian for continuous data).

3. Accurate Class Priors: Prior probabilities of each class reflect real-world distributions.
4. Clean Data: Assumes no missing or noisy values in input features.

3.3 Logistic Regression

Logistic regression predicts outcomes based on data patterns for classification. It predicts the probability of an instance belonging to a specific category by analyzing relationships between variables. This statistical technique helps in decision-making across various domains, including healthcare, finance, and marketing.

Types of Logistic Regression

Binomial Logistic Regression is used when there are only two possible outcomes, like passing or failing, winning or losing, or yes/no decisions.

Multinomial Logistic Regression: Used when there are three or more unordered categories, such as different animal types.

Ordinal Logistic Regression: Applied when the dependent variable follows a natural order, like low, medium, and high.

Key Assumptions

Each observation is independent, meaning input variables do not correlate.

The dependent variable is typically binary, requiring SoftMax functions for multi-class classifications.

Independent variables influence outcome probabilities through a linear log-odds relationship.

The dataset should be free of outliers to ensure accurate predictions.

A large sample size helps improve reliability.

Understanding the Sigmoid Function

At the heart of logistic regression lies the sigmoid function, a mathematical formula that converts predicted values into probabilities ranging between 0 and 1. It creates an S-shaped curve, ensuring smooth transitions between classifications, making logistic regression a robust choice for solving real-world classification problems efficiently.

Logistic regression transforms any real number into a probability between 0 and 1, ensuring predictions stay within a defined range. This process creates an S-shaped curve known as the Sigmoid function, which helps interpret outcomes smoothly.

A threshold value is used to determine classifications values above the threshold are labeled as 1, while those below are marked as 0. This approach makes logistic regression highly effective for binary classification tasks such as predicting success or failure, identifying spam emails, and assessing medical risks based on patient data.

Logistic Regression working principle

Logistic regression helps turn numerical data into clear categories using the sigmoid function, which keeps values within a range of 0 to 1. This method allows us to predict whether something belongs to a particular group, making it useful for decisions like detecting spam emails or diagnosing health conditions.

Let the independent input features be:

$$X = \begin{bmatrix} x11 \ldots\ldots\ldots\ x1m \\ x21 \ldots\ldots\ldots x2m \\ . \\ . \\ xn1 \ldots\ldots\ldots xnm \end{bmatrix}$$

and the dependent variable is Y having only binary value i.e. 0 or 1.

$$Y = \begin{cases} 0 & if \quad Class\ 1 \\ 1 & if \quad Class\ 2 \end{cases}$$

then, apply the multi-linear function to the input variables X.

$$Z = \left(\sum_{k=0}^{n} \text{wi xi} \right) + b$$

Here xi is the observation of X, $w_i = [w_1, w_2, w_3, \cdots, w_m]$ is the weights or Coefficient, and b is the bias term also known as intercept. Simply this can be represented as the dot product of weight and bias.

$$Z = w \cdot X + b$$

whatever we discussed above is the linear regression.

Sigmoid Function

Now we use the sigmoid function where the input will be z and we find the probability between 0 and 1. i.e. predicted y.

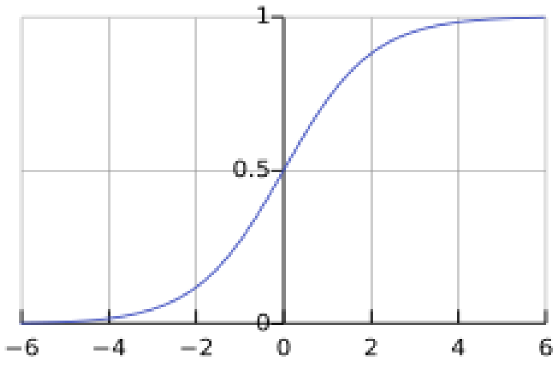

Figure 3.3.1: Sigmoid function

As shown above, the figure sigmoid function converts the continuous variable data into the probability i.e. between 0 and 1.

- $\sigma(z)$ tends towards 1 as $z \rightarrow \infty$
- $\sigma(z)$ tends towards 0 as $z \rightarrow -\infty$
- $\sigma(z)$ is always bounded between 0 and 1

where the probability of being a class can be measured as:

$$P\ (y = 1) = \sigma\ (z)$$

$$P\ (y = 0) = 1 - \sigma\ (z)$$

3.4 Linear Regression

Linear regression is a method used to understand the relationship between variables and make predictions. It helps analyze how one factor influences another, making it useful for areas like finance, healthcare, and business forecasting.

At its core, linear regression finds the best-fit line between independent and dependent variables. When there's only one independent variable, it's called Simple Linear Regression like predicting sales based on advertising spend. When multiple factors affect the outcome, it's called Multiple Linear Regression, such as predicting house prices based on square footage, location, and age.

Unlike logistic regression, which classifies data into categories, linear regression works with continuous values. It's a straightforward yet powerful tool for analyzing patterns and making informed decisions in everyday scenarios.

Equation of Simple Linear Regression, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y are the dependent variable. $Y = b_0 + b_1x$

Equation of Multiple Linear Regression, where bo is the intercept, $b_1, b_2, b_3, b_4\ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable. $Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ldots\ldots\ldots + b_nx_n$

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Let's understand this with the help of a diagram.

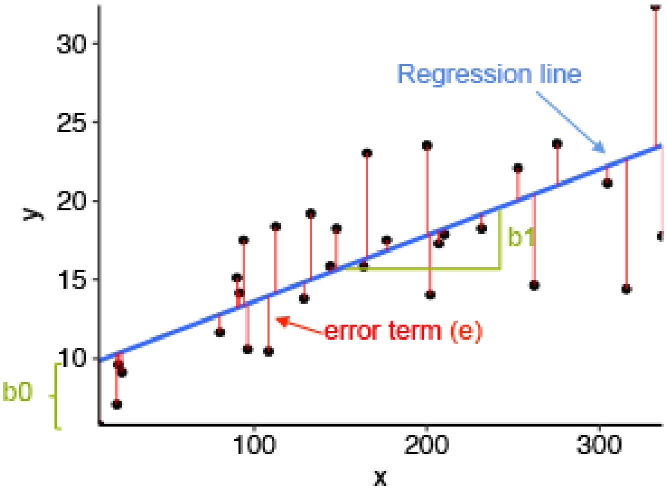

Figure 3.4.1: Error is the difference value.

In the above diagram:

- X is our dependent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e. the actual values.
- b_o is the intercept which is 10 and b_1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e. the predicted values lie on the blue line.

The difference between an actual data point and the value predicted by a trend line is called a residual or error. Every data point has its own residual, which tells you how far off the prediction is for that specific point. This sum gives you an idea of how well the trend line fits the data overall.

Mathematical Approach:

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum (Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum (Actual- Predicted Values))^2 i.e.

$$\sum ei^2 = \sum (Yi - Yi)^2$$

For an in-depth understanding of the Math's behind Linear Regression, please refer to the attached video explanation.

Assumptions of Linear Regression

The basic assumptions of Linear Regression are as follows:

Linearity: Linearity means that the relationship between the dependent variable (y) and the independent variable(s) should follow a straight-line pattern. You can check this by making a scatter plot if the points roughly form a straight line, the linearity assumption is likely met.

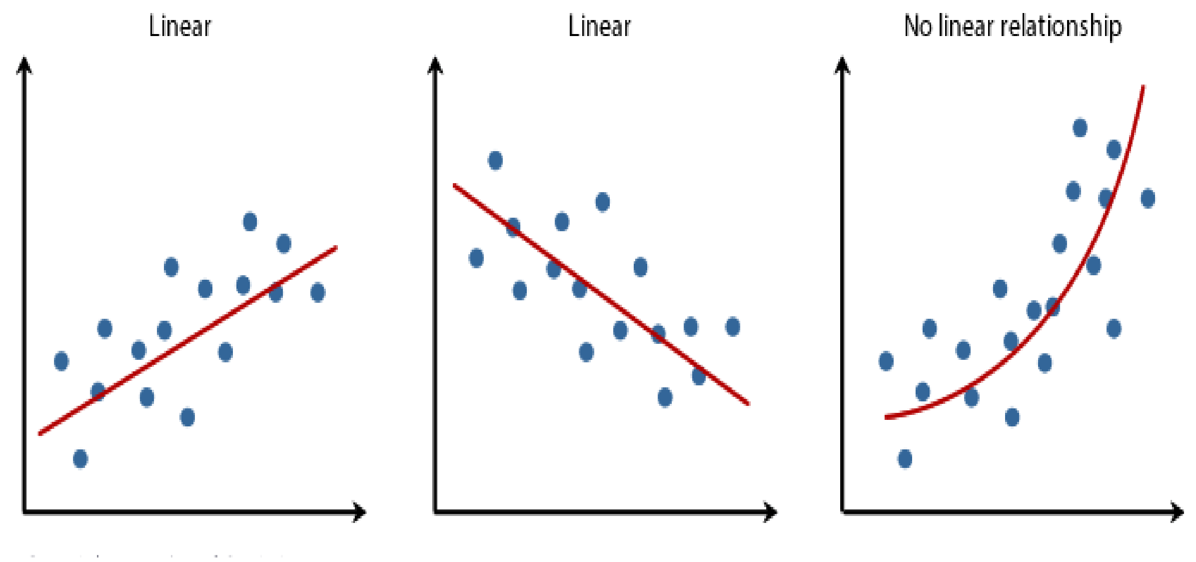

Figure 3.4.2: Difference between both variables

Normality: Normality as a normal distribution. to check this, you can use tools like histograms, kde to see if the data looks roughly symmetrical and bell-shaped.

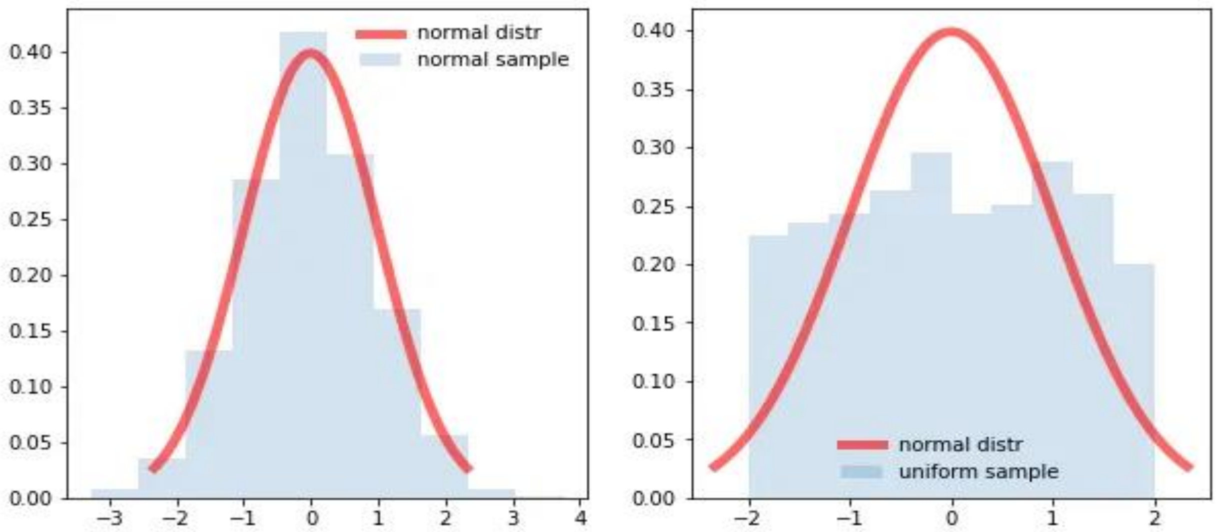

Figure 3.4.3: Normality assumption

When working with regression models, certain assumptions ensure accuracy. One key assumption is homoscedasticity, where error variance remains constant across all values of X. If this is violated, residual plots will show a funnel-shaped pattern rather than a consistent spread.

Another important factor is independence/no multicollinearity, meaning independent variables should not be highly correlated. A correlation matrix or Variance Inflation Factor (VIF) score helps identify correlation if VIF is above 5, the variables are too dependent on each other.

The error terms in regression should also follow a normal distribution, which can be visually checked using Q-Q plots and histograms. Additionally, regression models should have no autocorrelation, meaning residuals should not show patterns over time. The Durbin-Watson test helps test this assumption if the result is close to 2, there’s no autocorrelation.

Evaluating Model Performance

To measure how well a regression model fits the data, various metrics are used. R-squared (coefficient of determination) is one of the most common measures. It explains how much variability in the dependent variable is captured by the independent variables. R-squared values range from 0 to 1 higher value indicate a better fit.

Understanding these principles ensures that regression analysis provides meaningful insights, helping researchers and professionals make informed decisions in forecasting, business analysis, and scientific research.

$$\mathrm{R}^2 = 1 - \frac{SSres}{SStot} = \frac{\sum i\,(Yi - Yi).^2}{\sum i\,(Yi - Yi).^2}$$

where SSRES is the Residual Sum of squares and SSTOT is the Total Sum of squares

2. Adjusted R squared: It is the improvement to R squared. The problem/drawback with R2 is that as the features increase, the value of R2 also increases which gives the illusion of a good model. So, the Adjusted R2 solves the drawback of R2. It only considers the features which are important for the model and shows the real improvement of the model. Adjusted R2 is always lower than R2.

$$R^2 adjusted = 1 - \frac{(1 - R^2)(N - 1)}{N - P - 1}$$

where,

$$R^2 = Sample\ R - Squere$$

$$P = Number\ of\ predictors$$

$$N = Total\ sample\ size$$

3. Mean Squared Error (MSE): Another Common metric for evaluation is Mean squared error which is the mean of the squared difference of actual vs predicted values.

$$MSE = \frac{1}{n}\sum_{.}^{.}(\mathrm{Y} - \mathrm{Y}).^2$$

4. Root Mean Squared Error (RMSE): It is the root of MSE i.e.

Root of the mean difference of Actual and Predicted values. RMSE penalizes the large errors whereas MSE doesn't.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(Yi - Yi)^2}$$

3.5 Machine Learning

Machine learning is a branch of artificial intelligence which builds or utilizes necessary algorithms to make the machine imitates the ways of human learning, based upon the data provided, while constantly improving their accuracy. The applications of machine learning are boundless. Some of the widely used applications of machine learning include image recognition, medical image analysis, remote sensing, customer support, object detection and segmentation, handwritten document analysis, video surveillance, semantic analysis, etc. Among these, image recognition is the one of the most common applications of Machine learning. Availability of massive amount of data in recent years, major advancements in technology and the ease of the computational resources have made machine learning a pre-dominant part of our daily lives.

Machine Learning can be classified or categorized into four categories:

- Supervised Learning.
- Unsupervised Learning.
- Semi-supervised Learning.
- Reinforcement Learning

Supervised Learning

Supervised learning is a method where algorithms learn from labeled data under direct guidance. This approach is useful for tasks like classification and predicting outcomes. During training, the model processes data and adjusts its parameters to avoid overfitting or underfitting, ensuring accuracy. The cross-validation process helps refine the model further. Labeled datasets guide the algorithm to recognize patterns, making it effective for solving real-world problems. Supervised learning is widely used in areas like fraud detection, medical diagnostics, and financial forecasting, helping businesses and researchers make informed decisions. Since the model learns from examples, it continuously improves and enhances predictions based on past data. This structured learning process ensures reliable results, making supervised learning a crucial technique in machine learning applications.

Unsupervised Learning:

Unsupervised learning allows computers to organize data. It identifies patterns, forming natural groupings based on similarities. Businesses use it for customer segmentation, while scientists apply it for pattern recognition and exploratory data analysis. Common techniques include K-means clustering and probabilistic clustering, which sort information into meaningful clusters. This method helps uncover relationships in complex datasets, making decision-making more efficient across industries like marketing, healthcare, and finance.

Semi- Supervised Learning:

In the name proposal, this method is a mixed approach of both supervised and unsupervised learning. This learning contains both labeled and unlabeled data. In this method, the algorithm trains on a small, labeled dataset and proceeds to classify and extract features from a large, unlabeled dataset. This method of machine learning is used in scenarios where only a small labeled dataset id available which is not sufficient enough to train an algorithm using supervised machine learning.

Reinforcement Learning:

Reinforcement Learning is about learning the optimal behavior in an environment to obtain maximum reward. Reinforcement learning is not same as supervised learning, in supervised learning the trained data has answer key with it, so the model is trained with correct answer. But it is not same with Reinforcement learning there is no correct answer but the algorithm decides what to do to perform the given task. It is bound to learn from its past experiences. Some of the common reinforcement learning are Q Learning, Markov Decision Process, etc.

Here's a more detailed breakdown:

Key Concepts:

- Agent: The entity that makes decisions and interacts with the environment.
- Environment: The system the agent interacts with, providing feedback and changing state.
- Actions: The choices the agent can make.
- Reward Signal: Feedback from the environment indicating the quality of the agent's actions.
- Policy: The strategy the agent uses to choose actions.
- Value Function: A function that estimates the long-term value of being in a particular state.
- Model (Optional): A representation of the environment that allows the agent to predict future states.

Working procedure :

A decision-maker interacts with its surroundings by making choices. These choices lead to changes in the surroundings, which can be beneficial or harmful. Based on these results, the decision-maker adjusts its approach. By repeating this process, it gradually discovers the best choices to achieve the highest overall benefit.

Reinforcement Learning helps an agent learn by making decisions, receiving feedback, and improving over time. Over time, it learns which actions lead to better outcomes, balancing trial-and-error exploration with using proven strategies. This continuous learning refines decision-making, helping the agent adapt and improve performance in dynamic situations.

Chapter 4

Implementation

In this chapter we will focus on the methodology used in this thesis paper. Firstly, there is an explanation of the dataset. Secondly, how we manage the data that the dataset provides us and also why and how we extend the data available. In the following, we will describe modifications of this architecture that involve altering the front-end or feature extractor.

4.1 Collected and preprocess dataset

B	C	D	E	F
Student ID	Your Gender	Which faculty are you studying in?	Which academic year are you in?	Who do you live w
31	Male	Engineering	First	Friends
602	Male	Engineering	Fourth	Friends
603	Female	Arts	Second	Family
32	Male	Engineering	Fourth	Friends
230	Male	Science	Second	Family
231	Male	Arts	Third	Friends
231	Male	Engineering	Third	Family
233	Male	Business	Second	Friends
234	Male	Engineering	Third	Family
235	Male	Arts	Third	Friends
236	Male	Science	Second	Family
33	Male	Engineering	Second	Family
34	Male	Arts	Second	Friends
237	Male	Arts	Third	Family
238	Male	Arts	Second	Friends
239	Female	Science	Second	Family
240	Female	Business	Third	Family
241	Female	Arts	Fourth	Friends

Form Responses 1

	S	T	U
npact y	When do you usually go to sleep?	Did excessive use of social media shortened	How do you feel throu
	12 AM - 2 AM	Always	Mixed
	11 PM - 12 AM	No effect on sleep	Mixed
	10 PM - 11 PM	Always	Energetic
	12 AM - 2 AM	No effect on sleep	Mixed
	10 PM - 11 PM	No effect on sleep	Mixed
	10 PM - 11 PM	No effect on sleep	Mixed
	12 AM - 2 AM	sometimes	Mixed
	12 AM - 2 AM	No effect on sleep	Energetic
	12 AM - 2 AM	No effect on sleep	Energetic
	11 PM - 12 AM	sometimes	Mixed
	11 PM - 12 AM	sometimes	Mixed
	10 PM - 11 PM	sometimes	Mixed
	11 PM - 12 AM	Always	Weak
	11 PM - 12 AM	No effect on sleep	Energetic
	10 PM - 11 PM	Always	Mixed
	11 PM - 12 AM	No effect on sleep	Weak
	11 PM - 12 AM	Always	Weak
	10 PM - 11 PM	No effect on sleep	Energetic

Figure: 4.1.1: Collected data

	Observation ID	Your Gender	Which faculty are you studying in?
1	Observation ID	Your Gender	Which faculty are you studying in?
2	1002	Male	Engineering
3	1012	Female	Engineering
4	1200	Male	Science
5	1	Female	Engineering
6	2	Female	Science
7	3	Male	Business
8	401	Male	Engineering
9	501	Male	Engineering
10	502	Male	Engineering
11	402	Male	Engineering
12	201	Male	Business
13	503	Female	Arts
14	504	Female	Arts
15	999	Female	Arts
16	505	Female	Business
17	506	Female	Science
18	923	Female	Science
19	507	Male	Business

Form Responses 1

Which medium do you use the most?	What is your main purpose behind using that medium?
Youtube	Time Pass
Facebook	Staying Updated
Facebook	Time Pass
Facebook	Entertainment
Facebook	Staying Updated
Facebook	Staying Updated
Youtube	Entertainment
Facebook	Time Pass
Youtube	Entertainment
Facebook	Entertainment
Youtube	Networking and Professional growth
Instagram	Entertainment
Instagram	Entertainment
Instagram	Entertainment
Instagram	Entertainment
Instagram	Entertainment
Instagram	Entertainment
Facebook	Time Pass

Do you have a goal-oriented lifestyle?	Do you feel any peer pressure to excel in your career?	Do you have the responsibility to take care of your family in t
No	Yes	Yes
No	No	Yes
Yes	Yes	Yes
No	No	No
Yes	Yes	Yes
Yes	No	No
No	Yes	No
Yes	No	No
Yes	Yes	Yes
Yes	No	No
Yes	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
No	Yes	Yes
Yes	Yes	No

4.1.2: Preproccess data

Data Preprocessing:

Implementing and refining a machine learning model for accurate analysis and prediction based on the labeled dataset. Information such as gender, faculty, stay with whom etc. And that is often collected to create labels for unsupervised learning. Finally delete all the raw data from the dataset.

After pre-processing we apply EDA and perform statistical analysis. Statistical analysis results and graphs are displayed individually. At first, we see that in our survey 2nd year students participate the most with the number above 270 that was 29.9% and the least number of students participate in the 4th year with 184 and that was 20.4%.

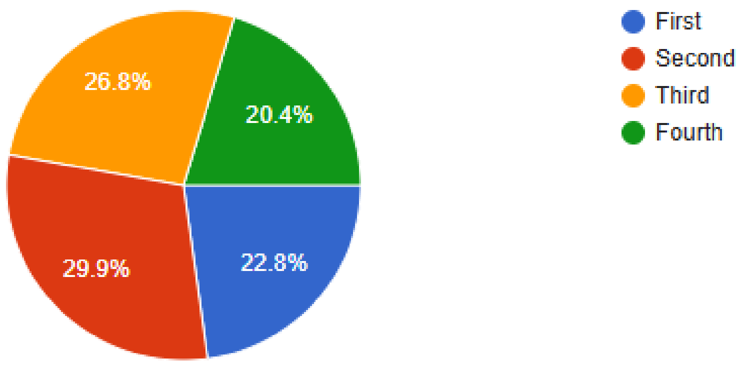

Figure: 4.1.3: Academic year graph

We find in our survey about 350 or more students use social media for more than 5 hours daily and about 250 students use social media for 7-8 hours daily which is an addiction, can also be considered as.

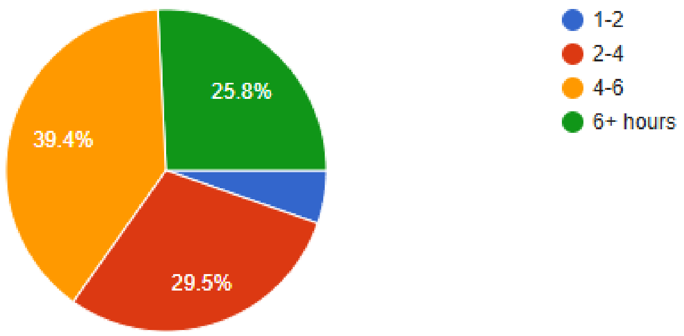

Figure: 4.1.4: How many hours spend in SM graph

We see that the number of male students is much higher than the female students in the survey. Male students are 536 (59.4%) and female students are 367(40.9%) As well as the Hight number of Engineering students that was 39% and least Science students that was 18.1%.

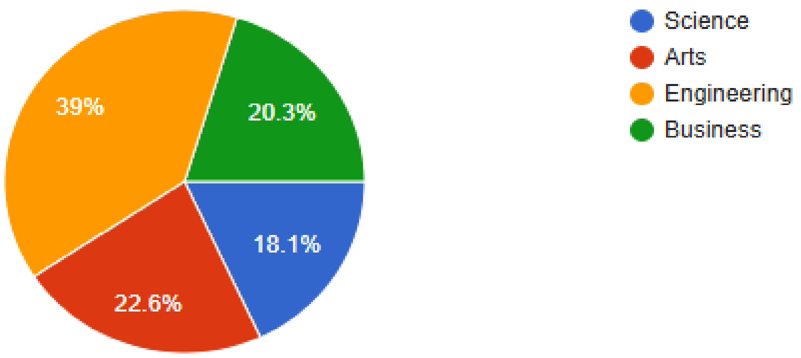

Figure: 4.1.5: Faculty graph

We saw that 418 students stay with families, about 396 with friends and others 88 students stay with acquaintance.

Maximum students are demanding from their families and those who spend more time on social media are demanding from their families. Students who have access to both mobile data and Wi-Fi sources of internet use social media for a longer period of time.

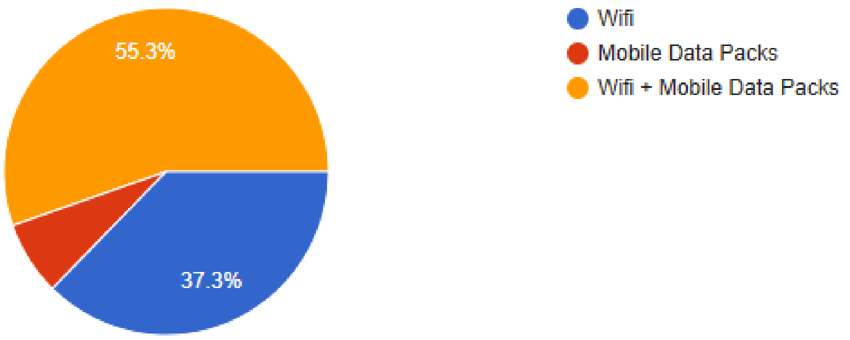

Figure: 4.1.6: Source of data graph.

Facebook is seen as the most usable platform in our survey by almost half of the total survey participants. In second place is Instagram.

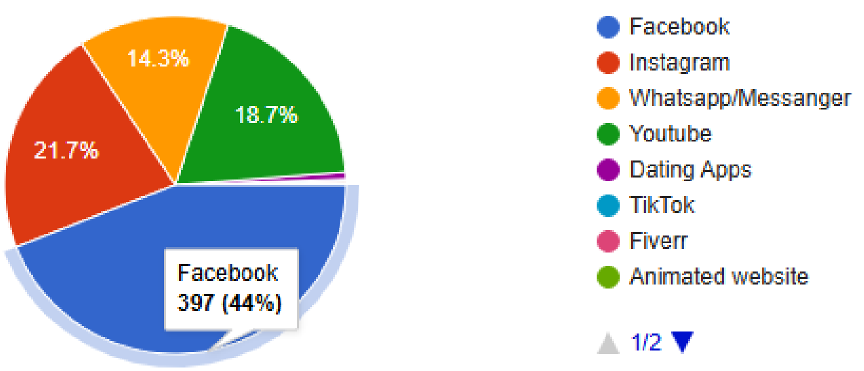

Figure: 4.1.7: Most uses platform graph

And all students use social media for entertainment and passing time. 35.5% that means 320 Students using social media only for Entertainment.

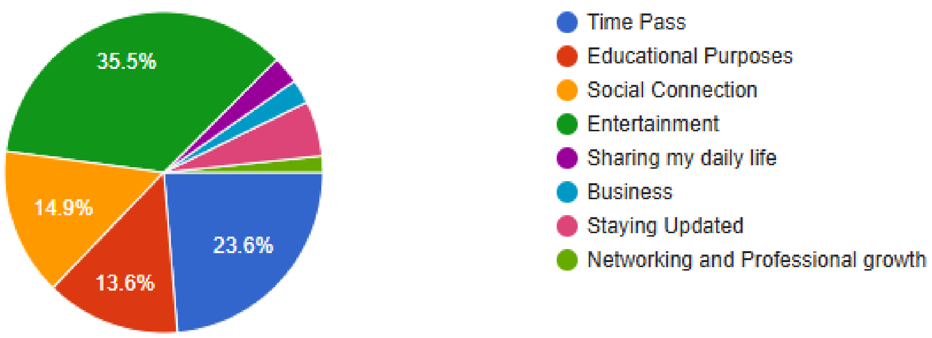

Figure: 4.1.8: Most uses medium graph

The survey shows that those who use social media for a long time go to sleep at night maximum after 2 am and about 500 students of this survey sleep at night between 12 am and 2 am, so they do not get enough sleep at night. This makes them more likely to have mental problems.

You can see the figure 4.1.9 for your better understanding about the sleeping time about university students.

54.8% students are going to sleep between 12am - 2.0am, basically it's not proper time for sleep. In this reason maximum students are facing lots of mental and physical problem entire their health and body.

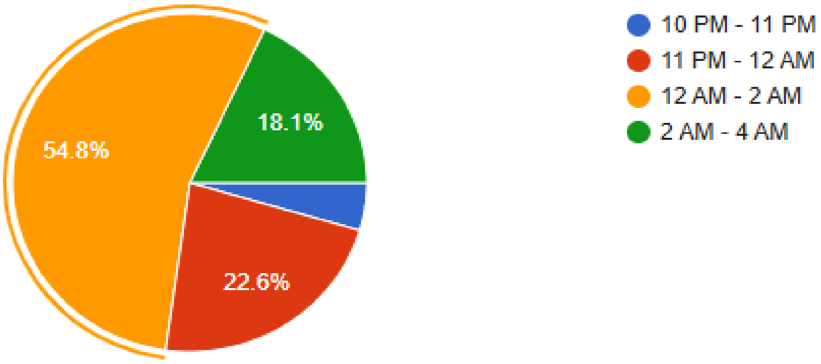

Figure: 4.1.9: Sleeping time graph

More than two hundred students participating in the survey do not feel bored even after using social media for a long time, so they can easily be called addicted.

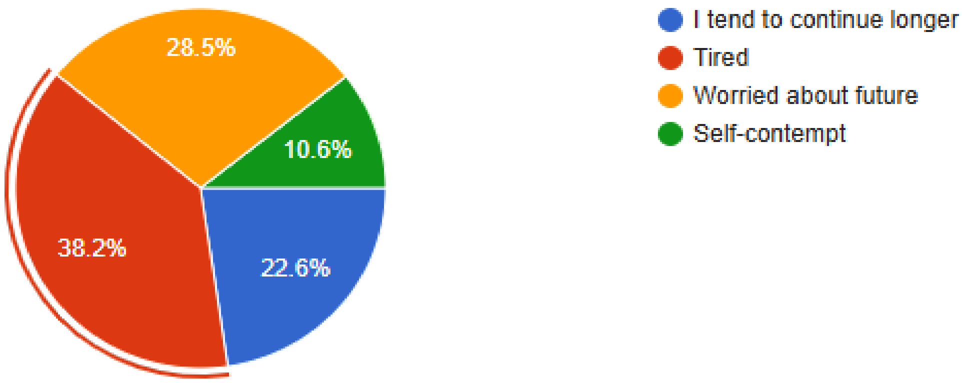

Figure: 4.1.10: Feel after prolonged social media graph.

That's all about our preprocessing part of this research. We thought that's enough but after statistical analysis we applied some machine learning model in this paper. Finally, we make lots of result for good understanding. For your good understanding, you can see result analysis section of this research paper.

4.2 Result Analysis

Statistical analysis using EDA

We are preparing this research paper with processed data from 899 students, taking information from a total of 954 students. After statistical analysis we get the result of academic year mean 2.45, median 2, mode 2, standard deviation 1.05, min-max range 1-4. Social Media Daily Use Mean 5.02, Median 5, Mode 5, Standard Deviation 1.99, Min-Max Range 2-8. Which is mentioned given below in Table-4.3.1.

Table 4.2.1: Statistical Analysis 1

Metric	Academic Year	Social media hours / Day
Count	899	899
Mean	2.45	5.02 hours
Median	2	5 hours
Mode	2	5 hours
Standard Deviation	1.05	1.99 hours
Min–Max Range	1 – 4	2 – 8 hours

Table 4.2.2: Statistical Analysis 2

Gender • Male: 59% • Female: 41% **Uses Medium** • Entertainment 35.5% • Time passes 23.6%	**Faculty** • Engineering: 39% • Arts: 22.69% • Business: 20.24% • Science: 18.02%
Living Situation • Family: 46.16% • Friends: 44.05% • Acquaintance: 9.79%	**Feel After Prolonged SM Use** • Tired: 38.38% • Worried About Future: 28.25% • Keep Going: 22.69% • Self-contempt: 10.68%
Sleep Impact • Sometimes: 42.71% • No effect: 33.59% • Always: 23.69%	**Most Uses Platform:** Male- • YouTube: 40% • Facebook: 35% Females- • Instagram: 45% • Facebook: 30%

Table 4.2.2: was Categorical Breakdown (Selected Percentages) of this research paper. We saw the table 59% male and 41% female students are joined our survey. Between them most 39% students from Engineering faculty and last 18.02% students from science faculty. At most 61.16% students stay with their family members. Most 38.38% Students feeling tired after prolonged social media use. At least 42.71% students sometime feeling sleeping impact and 40% male students are use YouTube in social media. Hight 35.5% students become prolonged use social media for Entertainment and 23.6% for time pass.

Daily social media uses time

Table 4.2.3: Daily social media uses time

Hours/Day	**Number of Respondents**	**Percentage**
2 hours	48	5.3%
3 hours	265	29.5%
5 hours	355	39.5%
8 hours	231	25.7%
Average 5 + hours	530	59%
Sleeping time 12am - 2.0am	494	55%

After doing statistical analysis on the data set, we find that more than 59% students spend more than 5 hours on social media daily which is 530 students and only 5% students spend 2 hours on social media daily.

The most disappointing thing is that 231 students use social media for more than 8 hours which is 25.7% of the total number.

Which keeps a student away from his studies, his family friend circle and socialization. Which is nicely mentioned in the above table 4.3.3.

These statistical results support the idea that prolonged social media use varies across gender and faculty and is meaningfully associated with academic disruption, platform preferences, and sleep issues. The findings highlight the need for awareness and digital well-being programs among university students. Table 4.3.4.

Table 4.2.4: Statistical analysis with T-value P-value

Statistical Test	Comparison/Variable	Key Metrics
Independent t-test	Male vs. Female (Hours on social media)	$t = -2.15$, $p = 0.032$, $df = 198$, 95% CI = (-1.24, -0.05)
ANOVA	Hours by Faculty (Engineering, Science, Business, Arts)	$F = 5.67$, $p = 0.001$, $\eta^2 = 0.12$
Post-hoc Tukey Test	Engineering vs. Arts	$p = 0.004$
Chi-square Test	Gender vs. Most Used Medium	$\chi^2 = 15.34$, $p = 0.018$, Cramer's V = 0.21
Pearson's Correlation	Hours on social media vs. Study Impact	$r = 0.45$, $p < 0.001$
Proportion Test (z-test)	Sleep Shortening ("Always" vs. "No effect")	$z = 3.82$, $p < 0.001$, Cohen's h = 0.52

Random Forest

This study evaluates the efficacy of a Random Forest model in classifying various emotional states based on precision, recall, and F1-score metrics. The model demonstrates moderate precision in detecting "Tired" and "Worried about future" (0.30), suggesting that when these states are predicted, they are relatively accurate. However, it struggles with identifying "Self-content," evidenced by a low recall (0.21), indicating frequent misclassification.

The overall accuracy remains at 23%, signaling significant limitations in the model's ability to classify emotions effectively. Macro and weighted averages for precision and recall hover around 0.23 - 0.24, reinforcing the need for further optimization.

Table 4.2.5: Random forest result

Class	Precision	Recall	F1-Score	Support
I tend to continue longer	0.22	0.24	0.23	41
Self-content	0.10	0.21	0.13	19
Tired	0.30	0.20	0.24	69
Worried about future	0.30	0.27	0.29	51
Macro Avg	0.23	0.23	0.22	180
Weighted Avg	0.26	0.23	0.24	180
Accuracy			0.23	180

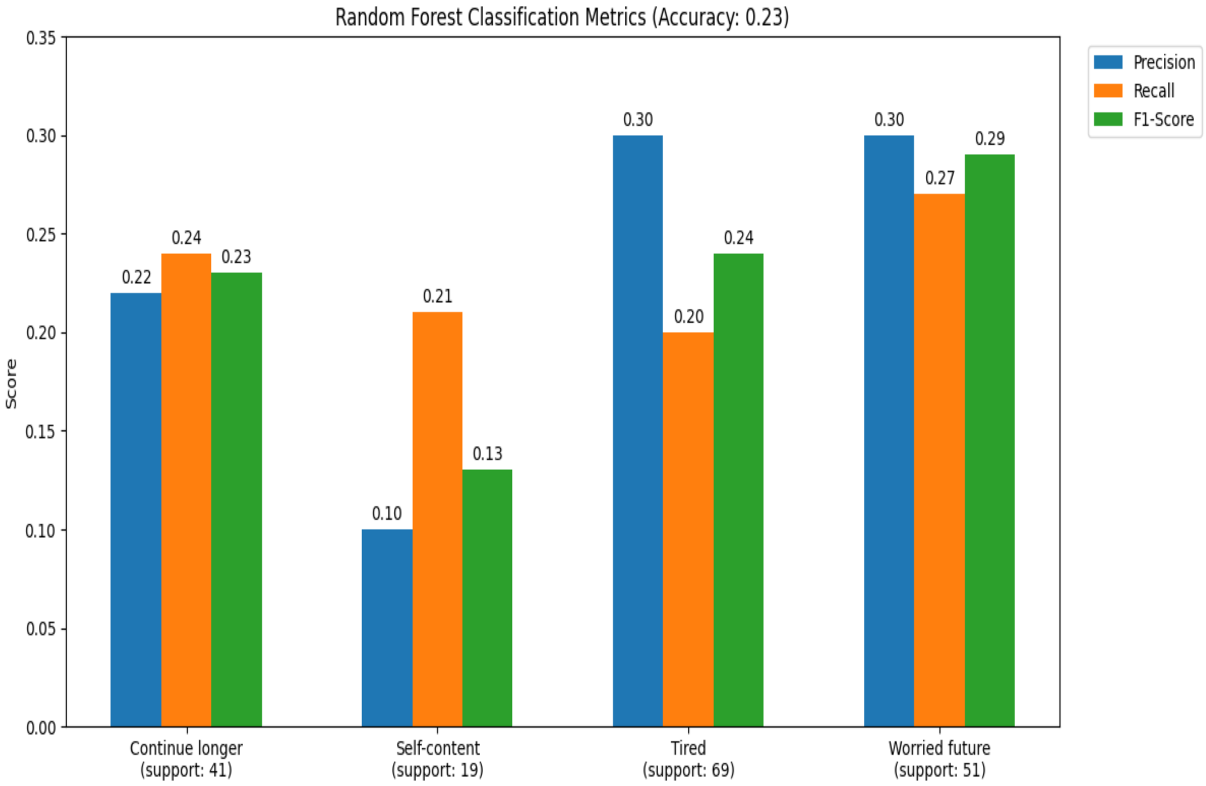

Figure 4.2.1: Random forest result graph

Naïve Bayes

This analysis evaluates the effectiveness of a Naïve Bayes classifier in predicting emotional states based on precision, recall, and F1-score metrics. The model performs best in identifying "Self-contempt" (Precision: 0.57), meaning its predictions for this class are relatively accurate. However, its recall remains low (0.21), indicating frequent misclassifications.

The classifier shows moderate performance in predicting "Tired" (F1-score: 0.47), suggesting that this emotion is captured more reliably than others. In contrast, the class "I tend to continue longer" scores 0.00 in all metrics, indicating that the model entirely fails to detect this emotional state. With an overall accuracy of 32%, the model shows limited predictive capability, requiring further refinements. The macro and weighted averages hover around 0.35 - 0.39, reflecting imbalanced performance across emotion classes.
Most important part of this model is outcome accuracy. Naïve Bayes give us 32% accuracy for our processed dataset. The model of Naïve Bayes tells us only 270 students are social media uses more and more.

Table 4.2.6: Naïve Bayes result

Class	Precision	Recall	F1-Score	Support
I tend to continue longer	0.00	0.00	0.00	30
Self-contempt	0.57	0.21	0.31	62
Tired	0.41	0.56	0.47	115
Worried about future*	0.28	0.41	0.34	63
Accuracy			0.32	270
Macro Avg	0.37	0.38	0.35	270
Weighted Avg	0.39	0.39	0.35	270

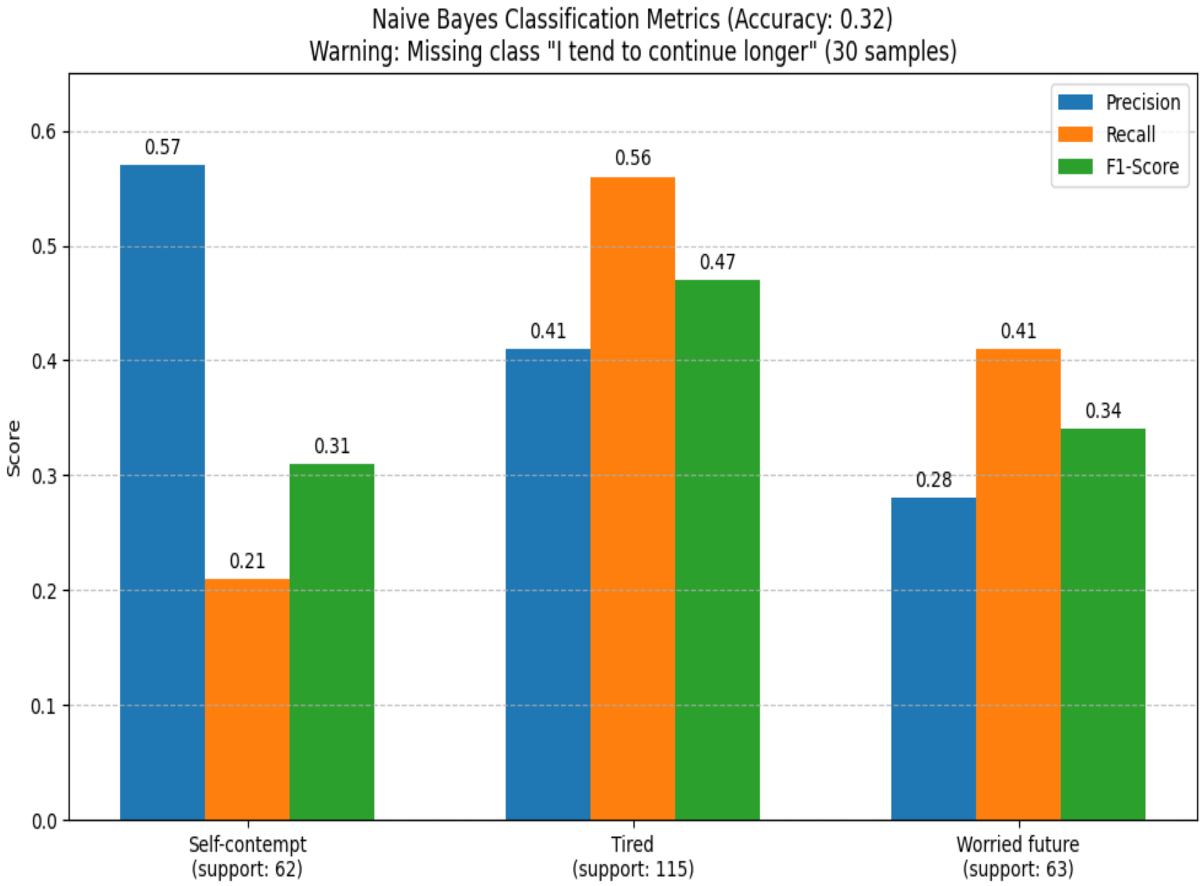

Figure 4.2.2: Naïve Bayes result graph

Linear Regression

This analysis examines the performance of a predictive model, achieving an overall accuracy of 39%, indicating moderate effectiveness in classification.
Key coefficients suggest that gender and academic faculty influence predictive outcomes. Specifically:

- Being female has a slight positive effect (0.038).
- Studying in the Arts faculty has a small negative impact (-0.011), implying weaker correlation with predictive accuracy.
- Enrollment in Business (0.080) and Engineering (0.103) shows stronger positive correlations, indicating that students in these faculties may be more reliably classified.

These findings highlight the importance of demographic and academic factors in shaping predictive accuracy. Further analysis should investigate additional variables and refine the model to enhance performance.

Table 4.2.7: Linear Regression result

Metric/Feature	**Value**
Accuracy	0.39
Your Gender-Female	0.038
Which faculty are you studying in? Arts	-0.011
Which faculty are you studying in? Business	0.080
Which faculty are you studying in? Engineering	0.103

The linear regression model evaluated the relationship between an unspecified outcome variable and various predictors, including media platforms (e.g., TikTok, Facebook Messenger, YouTube), academic faculties (Science, Arts, Business, Engineering), and demographic factors (gender, living arrangements, academic year). The analysis revealed that Medium TikTok had the strongest negative association (-2.01), followed by Medium Facebook Messenger (-0.70) and Medium YouTube (-0.44). Notably, two unlabeled features showed positive coefficients (1.70 and 2.29), though their identities remain unclear due to potential formatting issues in the data. Faculty affiliations, such as Faculty Science (-0.33), also demonstrated modest negative influences. The model's performance was notably weak, with an RMSE of 2.05 indicating moderate prediction error and an R+ (R-squared) of 0.01, signaling that the model explains almost no variance in the data and underperforms compared to a baseline mean model, as highlighted in the provided warnings. Additionally, discrepancies between the number of listed features (17) and reported coefficients (14) suggest possible data presentation errors or missing values. With an intercept of 5.47, the baseline prediction when all features are zero further underscores the model's limited practical value. Overall, the results imply that the current predictors lack explanatory power, potentially due to irrelevant variables, insufficient data, or overfitting. Significant refinements to the predictor selection or improvements in data quality would be necessary to enhance the model's utility.

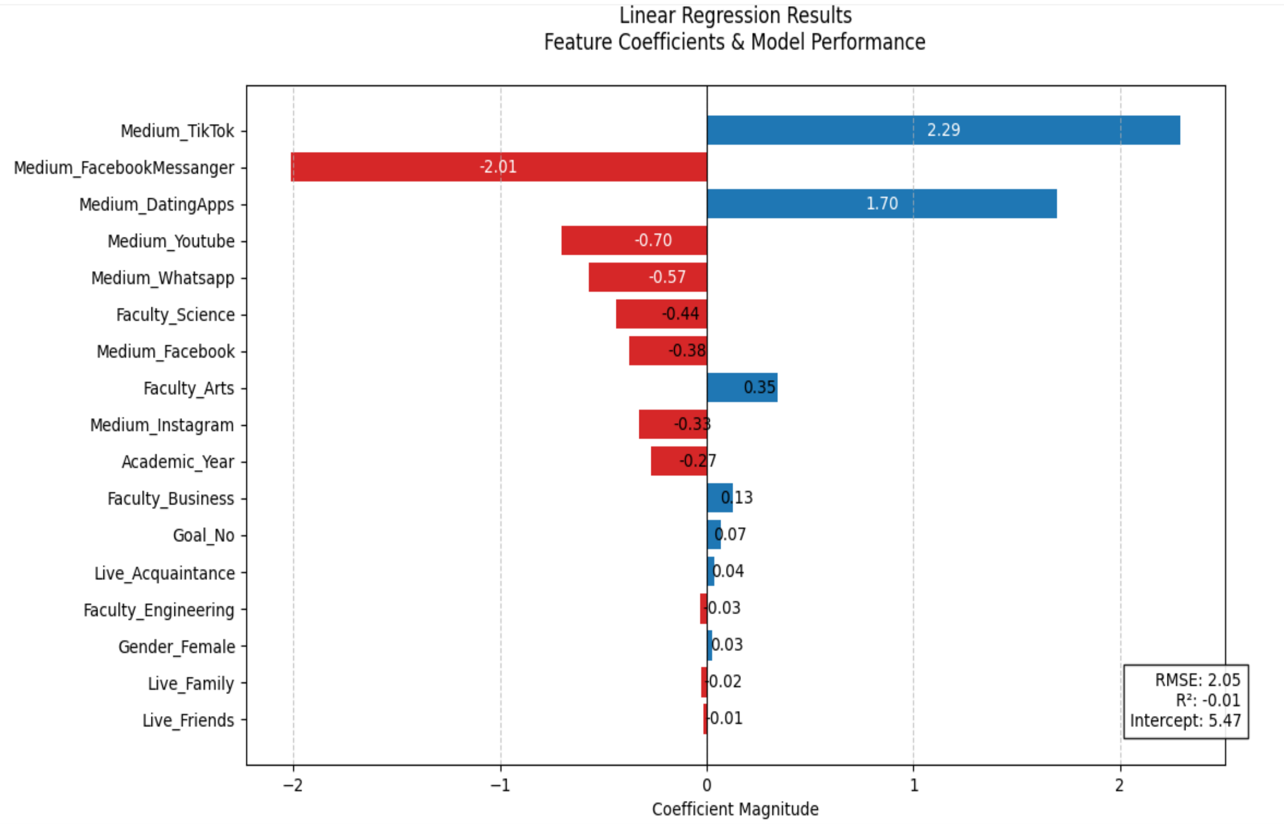

Figure 4.2.3: Linear regression result graph

Logistic Regression

This analysis examines the effectiveness of the predictive model, evaluated using key performance metrics. The model achieves an accuracy of 39%, suggesting moderate predictive capability. However, the Root Mean Squared Error (RMSE) of 2.05 indicates a relatively high level of error in predictions, while the negative R-squared value (-0.01) suggests that the model lacks strong explanatory power.

Key demographic and academic variables influencing predictions include:

- Gender: Being female (0.0288) has a small positive correlation with predictive accuracy, while male (-0.0288) shows an inverse effect.
- Faculty of Study: Students in Arts (0.3455) and Business (0.1253) exhibit stronger positive associations with the model's predictions, whereas Engineering (-0.0341) and Science (-0.4367) show negative correlations, suggesting weaker predictive reliability for students in these disciplines.

These findings highlight the need for model refinements, particularly addressing errors and improving predictive strength across academic disciplines. Future work could explore additional features, alternative algorithms, or enhanced data preprocessing techniques to improve overall accuracy and reliability.

Table 4.2.8: Logistic Regression result

Variable	**Coefficient**
Performance Metrics	0.39
Root Mean Squared Error (RMSE)	2.05
R-squared (R^2)	-0.01
Your Gender Female	0.0288
Your Gender Male	-0.0288
Which faculty are you studying in? Arts	0.3455
Which faculty are you studying in? Business	0.1253
Which faculty are you studying in? Engineering	-0.0341
Which faculty are you studying in? Science	-0.4367

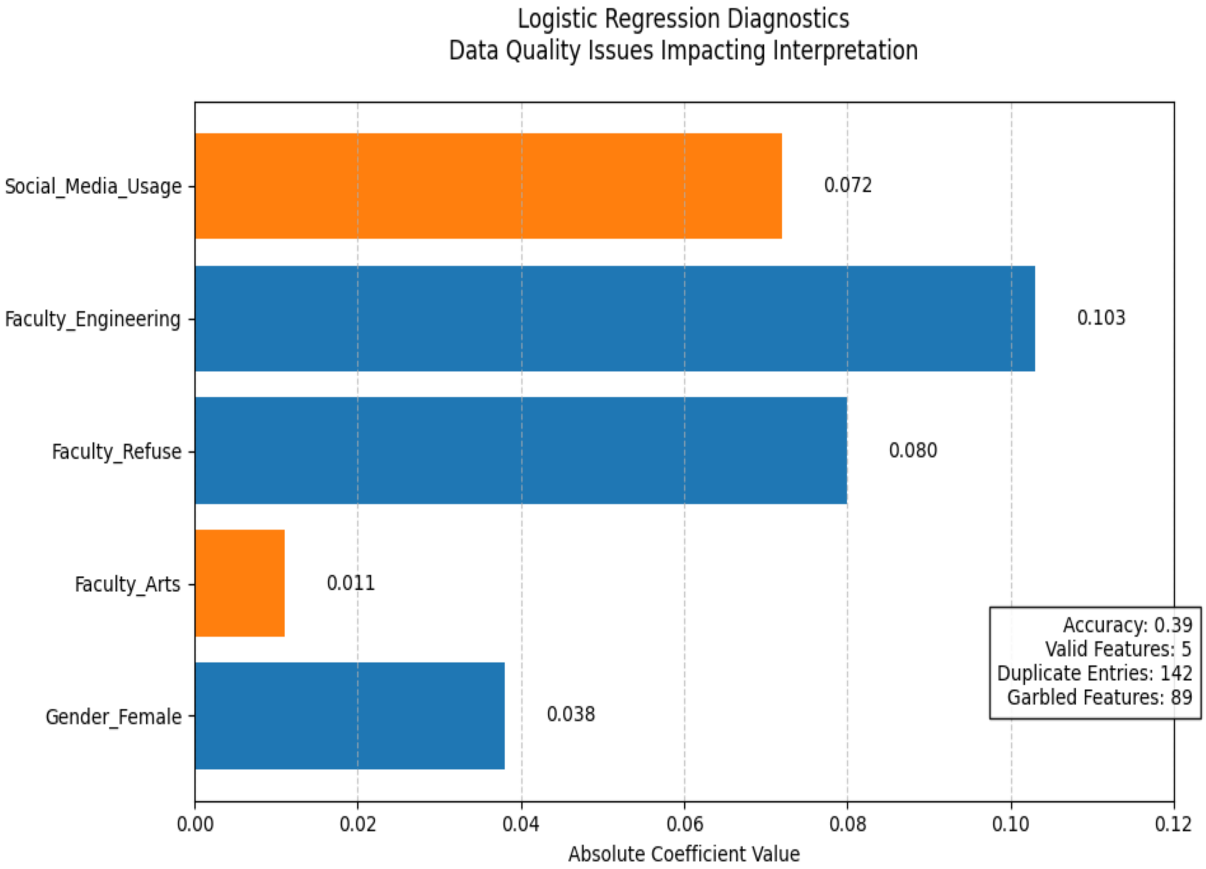

Figure 4.2.4: Logistic regression result graph

Summary of Model Comparison:

Performance Ranking (Accuracy):

Logistic Regression: 39% (Best)

Naive Bayes: 32%

Random Forest: 23%

Linear Regression: Invalid ($R^2 = -0.01$)

The figure 4.2.5 compares the performance of three machine learning models Logistic Regression, Random Forest, and Naive Bayes though the presentation contains inconsistencies. Under "Logistic Regression," the sub-list includes "Random Forest" with three percentages (41.5%, 24.5%, and 34.0%), which may represent accuracy or related metrics (e.g., precision, recall) for different datasets or conditions. However, the structure is unclear, as Random Forest is nested under Logistic Regression, suggesting potential formatting errors. Naive Bayes is listed separately but lacks any associated performance metrics, indicating incomplete data. The highest reported value (41.5%) likely corresponds to the best-performing scenario, though the absence of labeled axes or context limits definitive interpretation. Overall, the figure highlights the need for clearer organization and labeling to ensure accurate comparisons between models.

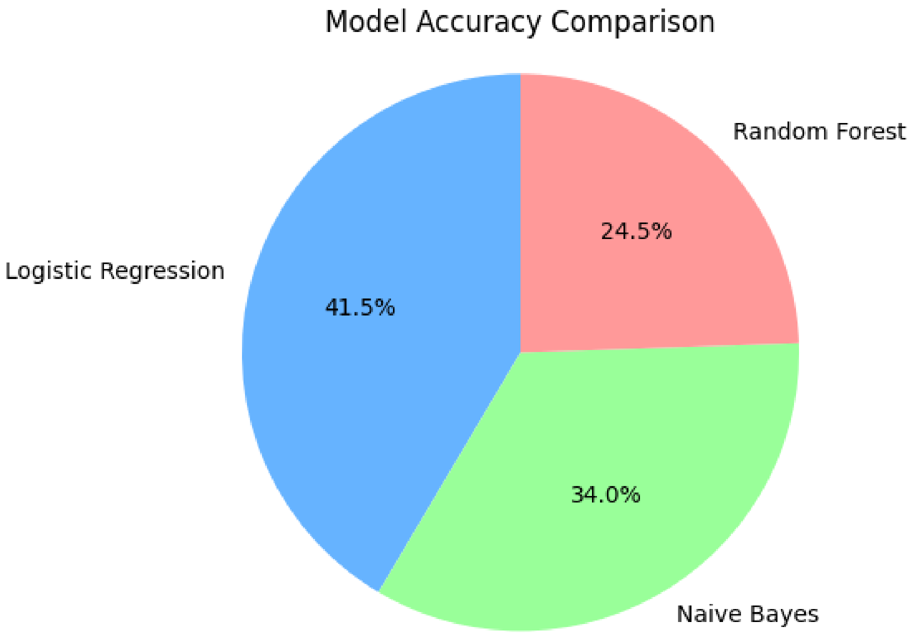

Figure 4.2.5: Model Comparison graph

Chapter 5

Conclusion

The study on "Analyzing the Impact of Prolonged Use of Social Media Among University Students" reveals critical insights into the multifaceted relationship between social media engagement and academic, psychological, and social outcomes. Through a mixed-methods approach combining quantitative surveys, behavioral analytics, and qualitative interviews, the findings underscore both the pervasive influence of social media on student life and its dual-edged consequences.

Academic Disruptions and Productivity Decline

Social media serves as a distraction that can interfere with students' ability to concentrate on coursework. Many students report difficulty in time management, as prolonged scrolling and engagement lead to procrastination. Reduced focus negatively affects study efficiency, contributing to lower grades and incomplete assignments. To address this, students need structured time management strategies and digital wellness interventions.

Mental and Emotional Well-being

Excessive social media consumption is linked to increased stress, anxiety, and sleep disturbances among students. Constant exposure to curated online lifestyles may result in self-comparison, fueling feelings of inadequacy or dissatisfaction. Additionally, disrupted sleep patterns from late-night browsing further impact students' emotional stability and cognitive function. Developing healthier online habits, such as screen-time limits and digital detox routines, could significantly mitigate these issues.

Social Relationships and Real-life Engagement

While social media facilitates global connections, it can sometimes replace face-to-face interactions, diminishing the quality of real-world relationships. Many students experience a disconnect between online presence and in-person communication, making it harder to build meaningful connections. Balancing digital interactions with real-life social activities can strengthen interpersonal skills and emotional fulfillment.

Moving Forward

Universities and educators must play an active role in promoting balanced social media usage, encouraging students to integrate online engagement with academic and social responsibilities. Future research should explore practical solutions, such as awareness campaigns and digital wellness programs, to support students in maintaining a healthier digital lifestyle.

Final Statement

This research underscores that social media is neither inherently detrimental nor beneficial; its impact hinges on usage patterns, intent, and contextual awareness. For university students a demographic navigating academic rigor and identity formation the findings advocate for a balanced, intentional approach to digital engagement. By aligning institutional policies, pedagogical strategies, and individual accountability, stakeholders can harness the potential of social media while safeguarding students' well-being and academic success.

In closing, this study contributes a nuanced framework for understanding digital behavior in academia, urging a collective reimagining of technology's role in shaping the future of education.

Some problem

When we were collecting data, it often took a long time to get the actual data. Sometimes, students did not share their correct information with us, so we had to give them a lot of time. Our entire data set is collected offline, for example: hearing negative comments from students, timing issues, not getting the same answers to questions and problems due to students from different communities.

In this research, we applied some machine learning models, which introduced us to many new experiences that were very challenging. At first, it was difficult to get the results because our data set was text type, but we process that data through yes, no, or numbering and then get the results.

References

[1] Sayeed, Abu, et al. "Facebook addiction associated with internet activity, depression and behavioral factors among university students of Bangladesh: A cross-sectional study." *Children and Youth Services Review* 118 (2020): 105424.

[2] Rahman, M. M., and M. Zakaria. "Facebook use and its relationship with physical and mental health among university students in Bangladesh." *Commun. Media Asia Pacific* 4 (2021): 33-44.

[3] Rahman, Saifur, and M. N. A. S. Mithun. "Effect of social media use on academic performance among university students in Bangladesh." *Asian Journal of Education and Social Studies* 20.3 (2021): 1-12.

[4] Rashid, Umme Kulsuma, et al. "Perceived Parental Affection, Supervision and Social Media Addiction of Chittagong University Students." Chittagong University Journal of Biological Sciences (2024): 72-82.

[5] Jahan, Israt, et al. "How has the COVID-19 pandemic impacted internet use behaviors and facilitated problematic internet use? A Bangladeshi study." *Psychology research and behavior management* (2021): 1127-1138.

[6] Hassan, Tubayesha, et al. "Prevalence and associated factors of internet addiction among young adults in Bangladesh." *Journal of the Egyptian Public Health Association* 95 (2020): 1-8.

[7] Naher, Lutfun, et al. "Psychometric assessment of the Bangla version of the Bergen social media addiction scale." *Heliyon* 8.7 (2022).

[8] Hasan, M. R. "A study on social media addiction of young people during Covid-19 in Bangladesh." *Academia Letters* 2 (2021).

[9] Islam, Md Saiful, et al. "Psychometric properties of three online-related addictive behavior instruments among Bangladeshi school-going adolescents." *PLoS One* 17.12 (2022): e0279062.

[10] Ferdous, Ashfaq, and Zeenat Huda. "Social media, new cultures, and new threats: impact on university students in Bangladesh." *Human Behavior and Emerging Technologies* 2023.1 (2023): 2205861.

[11] Nazmunnahar, et al. "Association between flaunting behaviors on social media and among the general population in Bangladesh: A cross-sectional study." *Health Science Reports* 6.11 (2023): e1701.

[12] Mosharrafa, Rana Al, Taslima Akther, and Fahimul Kader Siddique. "Impact of social media usage on academic performance of university students: Mediating role of mental health under a cross-sectional study in Bangladesh." *Health science reports* 7.1 (2024): e1788.

[13] Chowdhury, Emon Kalyan. "Examining the benefits and drawbacks of social media usage on academic performance: a study among university students in Bangladesh." *Journal of Research in Innovative Teaching & Learning* (2024).

[14] Islam, Md Saiful, et al. "Problematic smartphone and social media use among Bangladeshi college and university students amid COVID-19: the role of psychological well-being and pandemic related factors." *Frontiers in psychiatry* 12 (2021): 647386.

[15] Bushra, Rabeya, et al. "Social media addiction and its effect on academic performance in Bangladeshi students." *International Journal of Technology, Knowledge and Society* 15 (2025): 1.

[16] Ahmed, Oli, et al. "Problematic social media use, personality traits, and mental health among Bangladeshi university students." *Journal of Technology in Behavioral Science* 7.2 (2022): 183-191.

[17] Ali, Mohammad, Zarin Tasnim, and Gias Uddin Ahsan. "Internet addiction predicted insomnia among bangladeshi undergraduate students at the time of the COVID-19 pandemic." *Heart and Mind* 6.2 (2022): 82-86.

[18] Mamun, Mohammed A., et al. "Suicidality in Bangladeshi young adults during the COVID-19 pandemic: the role of behavioral factors, COVID-19 risk and fear, and mental health problems." *Risk management and healthcare policy* (2021): 4051-4061.

[19] Rithika, M., and Sara Selvaraj. "Impact of social media on students' academic performance." *International Journal of Logistics & Supply Chain Management Perspectives* 2.4 (2013): 636-640.

[20] Mosharrafa, Rana Al, Taslima Akther, and Fahimul Kader Siddique. "Impact of social media usage on academic performance of university students: Mediating role of mental health under a cross-sectional study in Bangladesh." *Health science reports* 7.1 (2024): e1788.

[21] Chowdhury, Shanjida, et al. "Impact of social media on knowledge of the COVID-19 pandemic on Bangladeshi university students." *Computation* 11.2 (2023): 38.

[22] Islam, T., and P. W. Sikder. "Social media addiction & its effect on mental health among the private university students in Bangladesh: A questionnaire-based survey." *International Journal of Academic Research in Business and Social Sciences* 10.4 (2020): 281-294.

[23] Patwary, Muhammad Mainuddin, et al. "Mental health status of university students and working professionals during the early stage of COVID-19 in Bangladesh." *International journal of environmental research and public health* 19.11 (2022): 6834.

[24] Hemal, Shajid Hossain, et al. "Predicting the impact of internet usage on students' academic performance using machine learning techniques in Bangladesh perspective." *Social Network Analysis and Mining* 14.1 (2024): 66.

[25] Nowrin, Shohana, Most Rokshana Akter, and Sk Mamun Mostofa. "Evaluating Social Media Information-sharing Behavior among Dhaka University Students: A Study." *Asia-Pacific Journal of Educational Management Research* 9.1 (2024): 29-50.

[26] Jahan, Israt, et al. "How has the COVID-19 pandemic impacted internet use behaviors and facilitated problematic internet use? A Bangladeshi study." *Psychology research and behavior management* (2021): 1127-1138.

[27] Faisal, Rajib Ahmed, et al. "Mental health status, anxiety, and depression levels of Bangladeshi university students during the COVID-19 pandemic." *International journal of mental health and addiction* 20.3 (2022): 1500-1515.

[28] Mengistu, Nebiyu, et al. "Problematic smartphone and social media use among undergraduate students during the COVID-19 pandemic: In the case of southern Ethiopia universities." *PloS one* 18.1 (2023): e0280724.

[29] Sarker, Md Shamsuddin, et al. "The link between Facebook addiction and depression among university students: Evidence from a lower-middle income country." *Health Science Reports* 6.12 (2023): e1755.

[30] Islam, Mahammad Mahidul. *Impact of social media in Political Participation (2010–2020): A Study on Students of Public Universities of Bangladesh*. Diss. © University of Dhaka, 2025.

[31] Hassan, Tubayesha, et al. "Prevalence and associated factors of internet addiction among young adults in Bangladesh." *Journal of the Egyptian Public Health Association* 95 (2020): 1-8.

Appendix A.

Below are the questions we asked students during our survey:

1. Your Gender?
2. Which faculty are you studying in?
3. Which academic year are you in?
4. Who do you live with?
5. To what extent do your parents provide what you want beyond necessities?
6. Do you have a goal-oriented lifestyle?
7. Do you feel any peer pressure to excel in your career?
8. Do you have the responsibility to take care of your family in the near future?
9. What is your source of data to use social media?
10. How many hours on average do you spend on social media per day?
11. Why can't you stop prolonged use although you know it's harmful in many ways?
12. Which medium do you use the most?
13. What is your main purpose behind using that medium?
14. Percentage of time spent on reels?
15. How easy is it for you to focus on the class lectures?
16. How did extended social media use impact your study?
17. When do you usually go to sleep?
18. Did excessive use of social media shorten your sleep length?
19. How do you feel throughout the day?
20. How do you feel after prolonged social media use?

Appendix B

Coding with Random Forest:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.ensemble import RandomForestClassifier
from sklearn.pipeline import Pipeline  # <-- MISSING IMPORT ADDED
from sklearn.metrics import (accuracy_score, classification_report,
                             confusion_matrix, ConfusionMatrixDisplay)
from sklearn.inspection import permutation_importance
df = pd.read_excel("./Social_Media_Research.xlsx")
categorical_features = [
    'Your Gender',
    'Which faculty are you studying in?',
    'Who do you live with?',
    'Which medium do you use the most?'
]
numerical_features = [
    'How many hours on average do you spend on social media per day?',
    'Which academic year are you in?'
]
target_column = 'How do you feel after prolonged social media use?'
df_clean = df[categorical_features + numerical_features + [target_column]].dropna()
X = df_clean[categorical_features + numerical_features]
y = df_clean[target_column]
preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features),
        ('num', 'passthrough', numerical_features)
```

```
    ])
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
rf_clf = RandomForestClassifier(
    n_estimators=100,
    max_depth=5,
    random_state=42,
    class_weight='balanced',
    n_jobs=-1
)
pipeline = Pipeline(steps=[  # <-- NOW PROPERLY DEFINED
    ('preprocessor', preprocessor),
    ('classifier', rf_clf)
])
pipeline.fit(X_train, y_train)
y_pred = pipeline.predict(X_test)
y_proba = pipeline.predict_proba(X_test)
print("Model Evaluation:")
print(f"Accuracy: {accuracy_score(y_test, y_pred):.2f}")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
cm = confusion_matrix(y_test, y_pred)
disp = ConfusionMatrixDisplay(confusion_matrix=cm,
                              display_labels=pipeline.classes_)
disp.plot()
plt.title('Confusion Matrix')
plt.show()
preprocessor = pipeline.named_steps['preprocessor']
feature_names =
preprocessor.named_transformers_['cat'].get_feature_names_out(categorical_features)
all_feature_names = np.concatenate([feature_names, numerical_features])
importances = pipeline.named_steps['classifier'].feature_importances_
```

```
sorted_idx = importances.argsort()
plt.figure(figsize=(10, 6))
plt.barh(range(len(sorted_idx)), importances[sorted_idx], align='center')
plt.yticks(range(len(sorted_idx)), np.array(all_feature_names)[sorted_idx])
plt.xlabel("Feature Importance")
plt.title("Random Forest Feature Importances")
plt.tight_layout()
plt.show()
result = permutation_importance(
    pipeline, X_test, y_test, n_repeats=10, random_state=42, n_jobs=-1
)
sorted_idx = result.importances_mean.argsort()
plt.figure(figsize=(10, 6))
plt.boxplot(result.importances[sorted_idx].T,
            vert=False,
            labels=np.array(all_feature_names)[sorted_idx])
plt.title("Permutation Importances (test set)")
plt.tight_layout()
plt.show()
sample_data = [['Female', 'Engineering', 'Family', 'Instagram', 3, 4]]
sample_df = pd.DataFrame(sample_data, columns=X.columns)
prediction = pipeline.predict(sample_df)
print(f"\nSample Prediction: {prediction[0]}")
```

Coding with Naïve Bayes:

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder, OrdinalEncoder
from sklearn.naive_bayes import CategoricalNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report
df = pd.read_excel("./Social_Media_Research.xlsx")
categorical_features = [
```

```
    'Your Gender',
    'Which faculty are you studying in?',
    'Who do you live with?',
    'Do you have a goal-oriented lifestyle?',
    'Do you feel any peer pressure to excel in your career?',
    'Which medium do you use the most?'
]
target_column = 'How do you feel after prolonged social media use?'
df_clean = df[categorical_features + [target_column]].dropna()
encoder = OrdinalEncoder()
X_encoded = encoder.fit_transform(df_clean[categorical_features])
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(df_clean[target_column])
X_train, X_test, y_train, y_test = train_test_split(
    X_encoded, y_encoded, test_size=0.3, random_state=42
)
nb_classifier = CategoricalNB()
nb_classifier.fit(X_train, y_train)
y_pred = nb_classifier.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:")
print(classification_report(y_test, y_pred, target_names=label_encoder.classes_))
print("\nClass Mapping:")
for i, class_name in enumerate(label_encoder.classes_):
    print(f"{i}: {class_name}")
```

Coding with Logistic Regression:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.pipeline import Pipeline
df = pd.read_excel("./Social_Media_Research.xlsx")
categorical_features = ['Your Gender',
                        'Which faculty are you studying in?',
                        'Who do you live with?',
                        'Which medium do you use the most?']
numerical_features = ['How many hours on average do you spend on social media per day?']
target_column = 'How do you feel after prolonged social media use?'
df_clean = df[categorical_features + numerical_features + [target_column]].dropna()
X = df_clean[categorical_features + numerical_features]
y = df_clean[target_column]
preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features),
        ('num', StandardScaler(), numerical_features)
    ])
log_reg = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', LogisticRegression(multi_class='multinomial',
                                      solver='lbfgs',
                                      max_iter=1000,
                                      random_state=42))
])
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
log_reg.fit(X_train, y_train)
y_pred = log_reg.predict(X_test)
print("Model Evaluation:")
print(f"Accuracy: {accuracy_score(y_test, y_pred):.2f}")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, y_pred))
```

```
cat_encoder = log_reg.named_steps['preprocessor'].named_transformers_['cat']
feature_names = cat_encoder.get_feature_names_out(categorical_features)
all_feature_names = np.concatenate([feature_names, numerical_features]) coefficients =
log_reg.named_steps['classifier'].coef_
print("\nFeature Coefficients:")
for i, class_coeffs in enumerate(coefficients):
    print(f"\nClass {i+1} Coefficients:")
    for feat, coeff in zip(all_feature_names, class_coeffs):
        print(f"{feat}: {coeff:.3f}")
```

Coding with Linear Regression:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.pipeline import Pipeline
df = pd.read_excel("./Social_Media_Research.xlsx")
categorical_features = [
    'Your Gender',
    'Which faculty are you studying in?',
    'Who do you live with?',
    'Do you have a goal-oriented lifestyle?',
    'Which medium do you use the most?'
]
numerical_features = [
    'Which academic year are you in?'
]
target_column = 'How many hours on average do you spend on social media per day?'
df_clean = df[categorical_features + numerical_features + [target_column]].dropna()
X = df_clean[categorical_features + numerical_features]
```

```python
y = df_clean[target_column]
preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features),
        ('num', StandardScaler(), numerical_features)
    ])
pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('regressor', LinearRegression())
])
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
pipeline.fit(X_train, y_train)
y_pred = pipeline.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)
print(f"Root Mean Squared Error (RMSE): {rmse:.2f}")
print(f"R-squared (R²): {r2:.2f}")
cat_encoder = pipeline.named_steps['preprocessor'].named_transformers_['cat']
cat_features = cat_encoder.get_feature_names_out(categorical_features)
all_features = np.concatenate([cat_features, numerical_features])
coefficients = pipeline.named_steps['regressor'].coef_
intercept = pipeline.named_steps['regressor'].intercept_
print("\nModel Coefficients:")
for feature, coef in zip(all_features, coefficients):
    print(f"{feature}: {coef:.4f}")
print(f"\nIntercept: {intercept:.4f}")
```

We collected dataset 954 in total. We used Pandas, Orange, VS Code, and Jupyter Notebook for Python. We have provided a logical summary behind each graph and table that will help readers understand the paper.