

**A MACHINE LEARNING APPROACH TO SMARTPHONE
ADDICTION PREDICTION**

BY

**NAHID AHMED NILOY
ID: 201-15-3584**

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

Supervised By

Mr. Shah Md Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Sabab Zulfiker
Senior Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project titled “A Machine Learning Approach to Smartphone Addiction Prediction”, submitted by Nahid Ahmed Niloy to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 24th January, 2024.

BOARD OF EXAMINERS



Dr. Md. Ismail Jabiullah

Professor

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

Chairman



Taslima Ferdous Shuva

Assistant Professor

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Sharun Akter Khushbu

Senior Lecturer

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Dr. Risala Tasin Khan

Professor

Institute of Information Technology

Jahangirnagar University

External Examiner

DECLARATION

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

Supervised by:

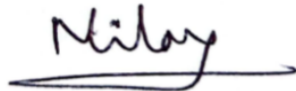


Mr. Shah Md Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Sabab Zulfiker
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:



Nahid Ahmed Niloy
ID: 201-15-3584
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

I am really grateful and wish my profound indebtedness to **Mr. Shah Md Tanvir Siddiquee, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of “*Data Science and Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

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

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

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

ABSTRACT

In our tech-filled world, smartphones have become a big part of our lives, but sometimes, people find themselves using them a bit too much, causing issues for themselves and those around them. This happens because we often seek quick rewards, spend too much time on social media, and feel a boost of happiness when using our phones. This research focuses on creating a machine learning model to detect and alert users about addictive smartphone habits. By fostering self-awareness, individuals may mitigate addiction's adverse effects, potentially leading to healthier lifestyles. Furthermore, increased awareness, especially among authorities, could preemptively curb addiction's onset. Drawing from 1203 responses across 26 questionnaires, data underwent meticulous handcrafted labeling and normalization for preprocessing. Testing various machine learning algorithms revealed random forest achieving a remarkable 97.51% accuracy, indicating substantial feature independence within the dataset. This model generates numerical scores or classifications, offering precise insights into addiction levels. The innovation lies in empowering individuals with information about their addictive tendencies, facilitating informed decisions about device usage. Proactive intervention against smartphone addiction holds promise in enhancing personal well-being and societal health. This predictive model's implementation could revolutionize addiction management, enabling early identification and intervention. By providing users with actionable insights, it aspires to not only curb addiction but also cultivate a healthier relationship with technology, fostering a balanced digital lifestyle for individuals and communities alike.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	PAGE
CHAPTER 1: INTRODUCTION	1-5
1.1. Introduction	1
1.2. Motivation	2
1.3. Objective	2
1.4. Rationale of the study	3
1.5. Research Questions	4
1.6. Expected Outcome	4
1.7. Project Management and Finance	4
1.8. Report Layout	5
Chapter 2: BACKGROUND STUDY	6-12
2.1. Preliminaries	6
2.2. Related Works	6
2.3. Comparative Analysis	9
2.4. Scope of Problem	11
2.5 Challenges	12

Chapter 3: RESEARCH METHODOLOGY	13-32
3.1 Research Subject and Instrumentation	13
3.2. Data Collection Procedure	13
3.3. Statistical Analysis	16
3.4 Proposed Methodology	22
3.5 Implementation Requirements	32
Chapter 4: EXPERIMENT RESULT AND DISCUSSION	33-45
4.1. Experiment Setup	33
4.2. Experiment Performance Evaluation Techniques	33
4.3. Experiment Result and Analysis	35
4.4. Discussion	44
Chapter 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	46-48
5.1. Impact on Society	46
5.2. Impact on Environment	46
5.3. Ethical Aspects	47
5.4. Sustainability Plans	47
Chapter 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	49-50
6.1. Summary of the Study	49
6.2. Conclusion	49

6.3. Implication for Future Study	50
REFERENCES	51-53
PLAGIARISM REPORT	54

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Gender Distribution	17
Figure 3.2: Age Distribution	17
Figure 3.3: Occupation	18
Figure 3.4: Educational Qualification	18
Figure 3.5: Number of Hours spent on smartphone	19
Figure 3.6: Miss Planned Work Due to Smartphone use	19
Figure 3.7: Hard time Concentrating	20
Figure 3.8: Feel pleasant or excited	20
Figure 3.9: Checking social media right after waking up	21
Figure 3.10: Number of addicted and not addicted users	21
Figure 3.11: Steps of proposed Methodology	23
Figure 3.12: Before Applying Standard Scaler	24
Figure 3.13: After Applying Standard Scaler	25
Figure 4.1: Confusion Matrix	34
Figure 4.2: Confusion matrix for LR	35
Figure 4.3: ROC Curve for LR	36
Figure 4.4: Confusion matrix for ADB	36
Figure 4.5: ROC curve for ADB	37
Figure 4.6: Confusion matrix for SVM	37
Figure 4.7: ROC curve for SVM	38

Figure 4.8: Confusion matrix for RF	38
Figure 4.9: ROC curve for RF	39
Figure 4.10: Confusion matrix for KNN	39
Figure 4.11: ROC curve for KNN	40
Figure 4.12: Confusion matrix for DT	40
Figure 4.13: ROC curve for DT	41
Figure 4.14: Accuracy of different algorithms	44

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Comparative Analysis of Related Research Works	10
Table 3.1: Questionnaires Used and Their Related Factors	15
Table 4.1: Comparison of Model Performance	42

Chapter 1

INTRODUCTION

1.1. Introduction:

Smartphone is an electronic device that is embedded with an integrated computer which basically originated from a telephone. The use of operating system, web browser and the ability to install different software mainly differentiates it from typical telephone. Over the past decade, the technological advancement of smartphone has been huge. Smartphones have become indispensable tools in today's interconnected world, serving a multitude of functions that touch nearly every aspect of our daily lives. Unlike telephone, smartphone comes with a variety of features like video call, messaging, internet access, listening to music, sending email, reading books online, playing video games with friends, accessing social media apps like Facebook, Instagram, Snapchat, Twitter and so on. Moreover, this device serves as powerful tools for creativity, with high-quality cameras allowing us to capture and share moments instantly, and various apps enabling photo and video editing. The smartphone's utility extends to navigation, as GPS and mapping applications help us navigate unfamiliar places. However, addiction to smartphone hampers a human being in many ways. Excessive smartphone use can lead to digital addiction, causing individuals to become overly reliant on their devices and reducing their ability to focus on real-world activities. This addiction can negatively impact productivity and relationships [1]. Also, the use of smartphones before bedtime, especially due to the emitted blue light and engaging content, can disrupt sleep patterns and lead to insomnia [2]. Moreover, continuous and prolonged use of smartphone can lead to many physical health issues like text neck, eye strain, musculoskeletal discomfort and so on. It can also contribute in building up stress, anxiety and depression. The reason I want to detect smartphone addiction is because these phones have become a big part of our lives, and sometimes we use them too much. This can cause problems like not sleeping well or missing out on real-life fun. Using these reasons for motivation, the primary objective of this work is to find out if someone is addicted to their smartphone using real-time data acquired by questionnaires and applying supervised machine learning and help them realize if they are

using smartphone too much. This research will help us figure out how bad the addiction is, and to step in early if it's getting really bad. By figuring out when people might be using their phones too much, I can help them use them in a better way. This means they can still enjoy their phones but without the bad stuff. I want to make sure people stay healthy and happy while using their smartphones.

1.2. Motivation:

In our current way of life, Smartphone is a valuable asset. It makes our daily life comfortable in many ways. In spite of that, most people are being addicted to everyday which is disrupting their life as well as the lives around them. This addiction can be caused by various things like Instant Gratification, social media and Peer Pressure, fear of missing out on something, Dopamine release and so on. It leaves a negative effect on one's physical health like eye strain, sleep disturbance, poor posture, negative effects on mental health like stress, anxiety and depression, decrease productivity, social and interpersonal problem, academic and work problems and so on. So, reducing smartphone addiction is crucial for individual as well as a societies well-being. Different methods can be adapted for reducing this addiction like using time tracking apps, setting clear goals, turning of notifications, finding different kind of alternatives etc. But among them, the most important this is self-awareness which means acknowledging the problem which is the first step to change. So, in this research, I will create a model using machine learning which will detect problematic smartphone usage and alert the user about their addictive behavior towards smartphone. If the user becomes self-aware, he might be able to overcome this addiction and lead a healthy life. Also, if police awareness is increased using this model, this addiction can be defeated before it even starts.

1.3. Objective:

In our demanding environment, smartphone addiction can be a serious issue. In order to prevent ourselves from getting addicted to smartphone, machine learning can be much of

help. Machine learning is helpful in providing individual advice and suggestion based on someone's smartphone usage. Machine learning can also predict the likelihood of developing smartphone addiction. It can analyze large amount of data and can track a lot of people's usages.

So, the main objective of this paper is to create a predictive machine learning model that will be able to predict addictive smartphone usage and alert that specific user. My main goal is to create user awareness by providing individuals with information about their potential addiction, it can empower them to make informed decisions about their device usage. Also, as this model is based on machine learning, it can provide a numerical score or classification which indicates the level of addiction that allows accurate understanding of the issue.

1.4. Rationale of the study:

Since smartphone addiction is becoming more common and has been shown to have negative effects on both individuals and society as a whole, we must take action to curb its growth. Additionally, not many studies on smartphone addiction have been done from a Bangladeshi perspective. Furthermore, in many situations, machine learning can provide results more quickly and accurately. For this reason, I'm interested in working with smartphone addiction and machine learning.

A sector of artificial intelligence which we call "machine learning" uses a range of statistical, probability-based, and optimization methods that allow computers to "learn" from past events and identify difficult-to-find patterns in massive, noisy, or complicated data sets. Machine learning techniques are applied in many different fields, such as detection and classification. Machine learning is applied to various fields such as cancer prediction, software defect detection and systemic review, and disease identification in dermatology, among others. These days, a wide range of risk prediction and detection techniques use machine learning. In the case of highly complex issues, machine learning approaches might play an additional role and offer a comparison to regression results.

Given the wide range of applications for machine learning, I decided that I should use it for my prediction work.

1.5. Research Questions:

- What are the right questions to ask to figure out smartphone addiction?
- How can I identify these smartphone addicts and what is the most efficient way of identifying them?
- How much data should I collect and where to collect them?
- what will be the number of features for the dataset and what will it look like?
- Which machine learning algorithms do I use and will they be compatible with my data?
- Do I use the popular machine learning algorithms or do I use new ones?

1.6. Expected Outcome:

Smartphone addiction is on the way of becoming a major issue in our modern world. But people are not aware of the damage it can do to individuals and society. The model that I have built is based in machine learning which is one of the best approaches in this scenario. It can predict the likeliness of addiction based on pattern from a huge dataset. It can also provide precise result in a short period of time. Also, using the suggested model, uncovering insights and trends in smartphone addiction is be possible which will be of great help in raising people's awareness.

1.7. Project Management and Finance:

At the start, I made a Google Form with questions based on the scale of smartphone addiction. I made arrangements to get responses to those questions from a range of sources, especially students. Thus, I got my required dataset. After acquiring the data, I need to use machine learning language to determine the extent of their smartphone addiction.

In this case, financial assistance is not required. The assistance of the public in data collection is the primary source of support for this work.

1.8. Report Layout:

This report consists of 6 chapters which are given below:

- In chapter 1, I outlined the overall structure of our research and divided this chapter into several smaller chapters such as Introduction, Motivation, Objective, Rational of the Study and Expected Outcome.
- In chapter 2, I covered the prior research on smartphone addiction, the Scope of the Problem and Challenges.
- In chapter 3, I mentioned how I collected data, the methodology of this work, statistical analysis and implementations.
- In chapter 4, I discussed the overall result of our study and created a comparative analysis between different algorithms.
- In chapter 5, I talked about Impact on Society, Environment, The Ethical Aspects and Sustainability Plan of this study.
- In chapter 6, I wrapped up this study by discussing the Summery, Limitations, Conclusion and Further Study of this Work.

Chapter 2

BACKGROUND STUDY

2.1. Preliminaries:

The widespread use of smartphones has radically changed communication, access of information, and engagement with the outside world, becoming a vital element of modern life. But with smartphones becoming so common, concerns about smartphone addiction have grown. This phenomenon can have serious effects on daily life, mental health, and wellbeing. Out of this concern, researchers have worked on this matter. In this chapter I will be discussing these works done by previous researchers. I will mention how they collected their data, how many data were used, their processed model and most importantly the algorithm they used to build the model and their result. Researchers used different algorithms like Random Forest, Support Vector Machine, Naïve Bayes, Logistic Regression, K Nearest Neighbor, Decision tree and so on. I will also compare their results in this section of this paper. Later on, will talk about the scope of the problem and lastly, I will be mentioning the challenges that I faced while working on this matter.

2.2. Related Works:

A Mazumdar, G Karak, S Sharma conducted a survey with well-designed questionnaires and took 115 undergraduate college students to complete their survey [3]. Then the collected data was trained to form a clustering-based machine learning model. After clustering, the result indicated that male lean towards to use smartphones more than females to use books and female have the highest possession count of phones for more than 12 hours whereas male has less than 6 hours.

Aljomaa, Suliman S., et al. also conducted similar kind of survey among 416 university students [4]. They used the comparative descriptive method to analyze the data and found out that 48% of the participants were smartphone addicts.

Arora, Anshika, et al. approached this problem with a different method [5]. They built an android application to collect real-time app usage data. Then they applied linear classification model like Support Vector Machine and Logistic Regression for the prediction. The Support Vector Machine achieved an accuracy of 81.33% and F1 score was 0.81 whereas Logistic Regression achieved an accuracy of 82.67% and F1 score was 0.827.

Shin, C., & Dey, A. K. also used the same method for collecting data [6]. They used both android phone app and interviewing people to collect data. For machine learning model, 3 classifiers, naïve Bayes, Support Vector Machine and Logistic Regression were used. Among these classifiers, SVM performed the best with 89.6% accuracy and 0.707 F-measure.

Lee, Juyeong, and Woosung Kim used the survey method where a total of 29712 people participated of different age [7]. To train the model, Random Forest, Decision tree and Xgboost were used. For Decision tree, average accuracy was 74.56%, high recall 0.81 and normal recall 0.67. For Random Forest, average accuracy was 82.59%, high recall 0.80 and normal recall 0.85. For Xboost, average accuracy was 80.77%, high recall 0.81 and normal recall 0.80.

KOKLU N., TASPINAR G., SULAK SA also used the same method where the data was collected from 1143 people through questionnaires [8]. For machine learning approach, they used Support Vector Machine. After applying this model, they obtained 92.4% accuracy and the value of the area under the ROC curve of the model was 0.993.

Giraldo-Jiménez, Claudia Fernanda, et al. conducted this survey on 1247 students at a private university in Colombia [9]. Deep learning techniques were used to make predictions on unseen data for better accuracy. 7 classifiers were used for training the data. Among them, Random Forest got the highest accuracy of 76.6%, SVM got the highest Sensitivity of 93.1, Decision tree got the highest Specificity of 50.1 and highest Precision of 79.2. For ROC score, Random Forest and SVM got the highest score of 0.73.

Bisen, Shilpa, and Yogesh Deshpande used 30 questionnaires on 100 students to conduct the survey [10]. For creating a sample, they used random sampling and purposive sampling method. The result of the survey was that male students has higher dependency as the mean

score for sub variables for Health, Communication and Shopping Apps was a little more than the female students although the mean score of female students using smartphone for education apps and entertainment apps was somewhat greater than the mean score of male engineering students.

Achal, Fairuz Tanzim, Mosammat Suraiya Ahmmed, and Tanjim Taharat Aurpa used machine learning techniques like SVM, Logistic Regression, KNN, Decision tree, Random Forest, AdaBoost, XGBoost, LightGBM to examine the data that they collected by survey [11]. They used 23 factors to predict smartphone addiction. Among the techniques, SVM got the highest accuracy which is 0.82 and Random Forest got lowest which is 0.60. They used chi-square feature to evaluate the performance of the models.

Masaru Tateno, Dai-Jin Kim, et al conducted research on smartphone addiction among Japanese college students [12]. They used the short version of smartphone addiction scale for this. They gathered a total of 573 data from college students where 180 were male and 393 were female. They asked them various questions. After applying the SAS-SV, they found out that 22.8% male and 28.0% female were addicted to smartphone.

Uz Zaman, Nur, Akther, Afroza, et al used activity recognition and app usage behavior to detect problematic smartphone use [13]. For app data collection, they took 30 participants and took the data after 7 days using the app. 30 participants were taken for questionnaires where 12 of them were male and 18 of them were female. Their analysis showed that 28% participants were addicted and their accuracy was 87% using different evaluation metrics like Rank Index and Fowlkes-Mallows Index.

Seong-Soo Cha and Bo-Kyung Seo worked with smartphone addiction in middle school students [14]. They took 1824 participants which are all middle school students. 51% of them were male and 49% of them were female. After applying multiple linear regression analysis, they found out that 563 or 30.9% students were addicted or in high-risk group and remaining 1261 or 69.1% students were not addicted or in normal group.

S Aggarwal, S Gupta, et al used usage patterns of preinstalled applications and questionnaires to predict smartphone addiction [15]. They installed an activity tracker that they developed in the participants smartphones. They also used SAS-SV for questionnaires.

A total of 96 people, 47 male and 49 female, was selected. They used algorithms like decision tree, decision tree with ada boost, logistic regression, k nearest neighbor and naïve bayes as machine learning algorithms. After the experiment, decision tree and decision tree with ada boost got the highest accuracy of 75%. Also, they concluded the, the participants spent their most time in social media apps.

One of the major ways of addicting to smartphone is social media. Islam, M. Z., Jannat, et al researched-on Facebook addiction which is a popular social media app [16]. They used more than 1000 data from of different age, occupation and background. They used machine learning algorithms like support vector machine, k nearest neighbor, decision tree, naïve bayes, logistic regression, and random forest. They also used techniques like PCA to reduce data mathematically. The result showed that support vector machine got the highest accuracy of 85%.

Eichenberg, C., Schott, M., & Schroiff, A. worked with problematic smartphone use and compared the personality of students with and without problematic smartphone use [17]. A total of 497 data of students from different background were used in their work. They used materials like SAS, 10-item BFI, short version of BSI, and short version of social support questionnaires. After calculating the mean and standard deviation score, they figured out that, for personality, participants scored the best on openness and conscientiousness and the worst on neuroticism. As for problematic smartphone use, 15.1% of participants were the victim of problematic smartphone use.

2.3. Comparative Analysis:

Prediction and detection using machine learning algorithms and data mining techniques have already been the subject of some previous research. These days, stress prediction, game addiction detection, and other types of addiction detection have become more common uses for machine learning technology. This part presents a comparison of these related works. A comparison of various research studies with respect to their subject, methodology, and results is given in the table below:

TABLE 2.1: COMPERATIVE ANALYSIS OF RELATED RESEARCH WORKS

SL	Name of the Author	Description	Methodology	Result
1	Arora, Anshika, et al	Intelligent Model for Smartphone Addiction Assessment	Support Vector Machine, Logistic Regression	SVM accuracy: 81.33% LR accuracy: 82.67%
2	Shin, C., & Dey, A. K.	Automatically Detecting Problematic Use of Smartphones	Naïve Bayes, Support Vector Machine and Logistic Regression	Best performing algorithm: SVM Accuracy: 89.9%
3	Lee, Juyeong, and Woosung Kim	Prediction of Problematic Smartphone Use	Random Forest, Decision tree and Xgboost	Best performing algorithm: Random Forest Accuracy: 82.59%
4	TASPINAR, Gulsen, et al	Predicting Smartphone Addiction using Support Vector Machine	Support Vector Machine	SVM accuracy: 92.4% AUC score: 0.993
5	Giraldo-Jiménez, Claudia Fernanda, et al.	Smartphone's dependency risk analysis using machine-learning predictive models	Random forest, Decision tree, SVM	Best performing algorithm: Random Forest Accuracy: 76.6% SVM with the highest Sensitivity of 93.1

6	Achal, Fairuz Tanzim, et al	Severity Detection of Problematic Smartphone Usage	SVM, Logistic Regression, KNN, Decision tree, Random Forest, AdaBoost, XGBoost, LightGBM	Best performing algorithm: SVM Accuracy: 82%
7	Uz Zaman, Nur, Akther, Afroza, et al	Smartphone addiction detection using activity recognition and app usage behavior	Different evaluation metrics like Rank Index, Fowlkes-Mallows Index	Highest accuracy: 87%
8	S Aggarwal, S Gupta, et al	Smartphone addiction detection using activity tracker	Decision tree, Decision tree with Ada boost, Logistic Regression, KNN and Naïve Bayes	Best performing algorithm: Decision Tree and Decision Tree with AdaBoost Accuracy: 75%

2.4. Scope of Problem:

This research focuses on designing a model through data analysis and the application of machine-learning algorithms. This proposed model can predict the risk of smartphone addiction. This prediction will have a significant social impact.

The younger generation can reduce their dependency on smartphones. This model can be used by government health departments and educational institutions for a variety of tasks. It is dangerous to become addicted to excessive smartphone use and anyone can become addicted to it. So, this model could help ordinary people and responsible individuals reduce their use of smartphones. Parents worry too much about their children and the future, which might hinder a generation's development and reflect negatively on society. Also, this model

can identify individuals who are at risk of addiction. As a result, it can intervene early and alert that individual. Moreover, machine learning and artificial intelligence have been used for various object detection and disease prediction and identification, with promising results. In context of smartphone addiction, Machine learning can identify patterns easily that contributes to addiction. That's why I decided to develop a model of addiction risk prediction using machine learning.

2.5 Challenges:

No research is without challenges. Every research comes with a set of challenges. I also faced some challenges while conducting this research. First of all, the data collection phase. In this research, I used raw data for better relevance. But collecting raw data was tough. Most of the data was collected through google form. I double-checked to make sure that the form was engaging and captivating. Also. I had to make sure not to include too many questions which will bore the participant and as a result he might give random answers. Moreover, not many people were interested in providing their insight in this matter. After the data collection process, I faced some challenges in preprocessing part. To label the data, I had to use a threshold value. Selecting the correct value was challenging as there was no fixed value and each value related in different accuracy. Besides, some responses were not accurate so I had to find them and exclude them from the dataset. I also had some problems in implementing the algorithms but with the help of my supervisor, I manages to overcome the problem and continue with my work with enthusiasm.

Chapter 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation:

I used a variety of algorithms and composite models to extract the most accuracy from the dataset. As we already know Machine learning algorithms are universally acknowledged and popular for any type of prediction, identification and classification. I experimented with multiple machine-learning algorithms on my collected dataset to identify which algorithm would be best to satisfy my expectations and perform the best. I used different machine learning algorithms like Logistic Regression, Support Vector Machine, Random Forest, AdaBoost Classifier and K Nearest Neighbor. For programming language, I used 'Python' as it is widely used by researchers because of its popularity and simplicity. For datamining platform, I used 'Anaconda Navigator' and 'Spyder' along with 'Microsoft Excel' as the dataset. Also, the dataset was in CSV format for compatibility.

3.2. Data Collection Procedure:

To conduct this research, I worked with raw data. For data collection, I used survey-based questionnaires. But selecting the appropriate questions for this research was a tough choice. After going through different websites and articles, I came across a scale known as Smartphone Addiction Scale or SAS [18]. It is based on 33 questions based on 6 features where the internal-consistency test result (Cronbach's alpha) is 0.967. So, based on this scale, I chose 19 questions for this research from the questionnaires. These questions were based on the highest expert approval rating from the SAS where 7 experts gave their approval if the questions were relevant to this topic or not. Along with some other questions, I selected a total of 26 questions based on 6 factors for this research. The factors are as follows:

- Daily-life disturbance
- Positive anticipation

- Withdrawal
- Cyberspace-oriented relationship
- Overuse
- Tolerance

After selecting the questions, I created a google form to collect data from people. I used a six-point Likert scale to collect the data that was used for this research. Basically, it is a psychometric measurement scale that is frequently used in research and surveys to evaluate participants' attitudes, opinions, and feelings. It gives participants a variety of ways to indicate whether they agree, are satisfied, or disagree with a statement. Participants are given six response options on a six-point Likert scale, each of which represents a different degree of agreement or disagreement, satisfaction or dissatisfaction, frequency, or intensity. Usually, the scale goes from extremely satisfied to extremely dissatisfied, or from strongly agree to strongly disagree. The points are given below:

1. Strongly Disagree
2. Disagree
3. Somewhat Disagree
4. Somewhat Agree
5. Agree
6. Strongly Agree

After creating the form, I spread it using social media and survey-based platform. Most of the responses it got was from students especially university students. Collecting data was also a hard process as I was lacking manpower. Fortunately, I managed to get 1203 responses. From these 1203 responses, 991 were male and 212 were female. Also, I managed to collect data from different group of people based on various measures like age, occupation and educational qualification. But majority of the responses were from younger people mostly students. The questions that were used in the research are given below:

TABLE 3.1: QUESTIONARIES USED AND THEIR RELATED FACTORS

SL	Question	Number of Expert Approval	Factor
1.	Missed work due to smartphone	7	Daily-life Disturbance
2.	Hard time concentrating in work	7	
3.	Experience lightheadedness or blurred vision	5	
4.	Feel pain in the wrists or at the back of the neck	6	
5.	Feel pleasant or excited	4	Positive Anticipation
6.	Having fun using smartphone	3	
7.	Life would be empty without smartphone	4	
8.	Won't be able to stand not having a smartphone	7	Withdrawal
9.	Feel impatient and fretful when not holding smartphone	7	
10.	Never give up smartphone	6	
11.	Relationship strength of smartphone buddies	5	Cyberspace-oriented Relationship
12.	Use smartphone after waking up	5	
13.	Check smartphone not to miss conversation	6	
14.	Use smartphone longer than anticipated	7	Overuse
15.	Urge to use smartphone after using it	4	
16.	Prefer searching from smartphone than asking other people	1	
17.	Thought of shortening use time	4	Tolerance
18.	Failing at shortening use time	5	

19.	People calling out excessive smartphone use	6	
20.	Daily smartphone use (Hour)	3	

After collecting the data, I moved to data preprocessing stage to reshape this data for further processing to get the best result.

3.3. Statistical Analysis:

Most of the data that was used in this research was collected through google form. I selected those questions that I thought was best for this research. I collected a total of 1203 data where 767 people were addicted and 436 were not addicted. I have collected both demographic as well as related data. Most of the related data were scale based where the scale value was from 1 to 6. The major features of my data include missing work due to smartphone use, hard time concentrating due to overuse, felt pleasant or excited, having fun using smartphone. The dataset also has physical problems caused due to excessive smartphone use like felling pain in the wrists or at the back of the neck or experiencing lightheadedness or blurred vision while using smartphone. There are also features like checking smartphone after waking up, using smartphone not to miss any conversation, using smartphone more than anticipated and so on. These features help significantly in identifying problematic smartphone use. As for demographic data, I collected age, gender, occupation, academic qualification and time of smartphone use daily in hours. The visual representation of the demographic data is given below:

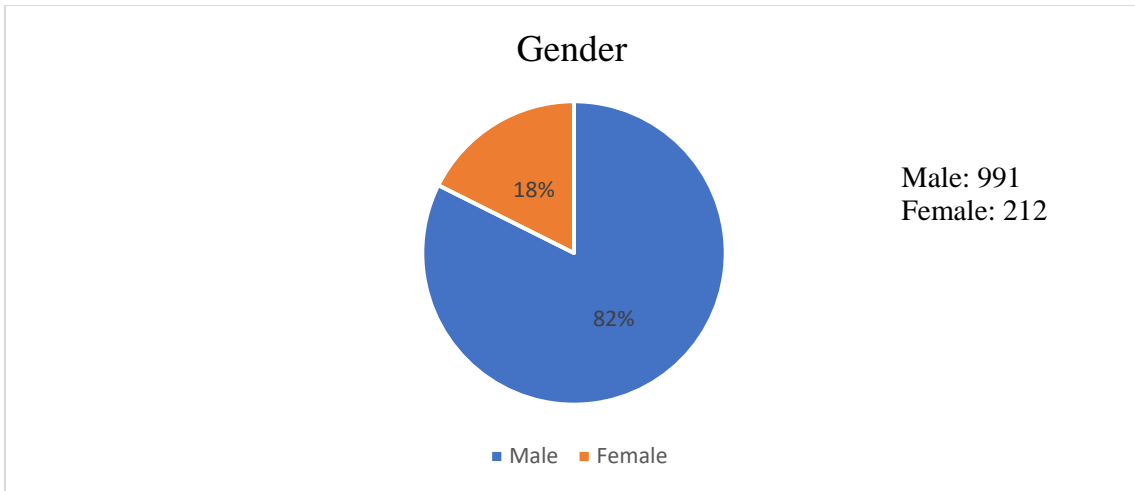


Figure 3.1: Gender Distribution

In the pie chart from figure 3.1, we can see that majority of the individuals who participated in this survey are male which is 82% and for the female participants, their number is 18%.

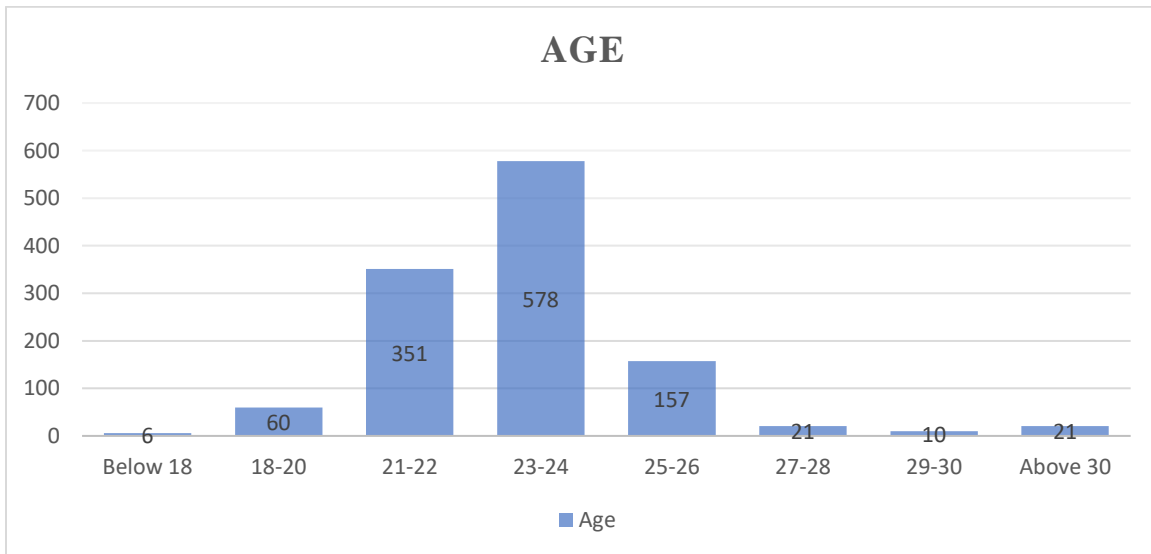


Figure 3.2: Age Distribution

In the figure 3.2, we can see that, most of the participants are from age group 23-24 where their number is 578. Age group 21-22 has the second highest number which is 351. If we go beyond age 26, the number significantly decreases. So, it is safe to say that this research was done mainly based on younger people.

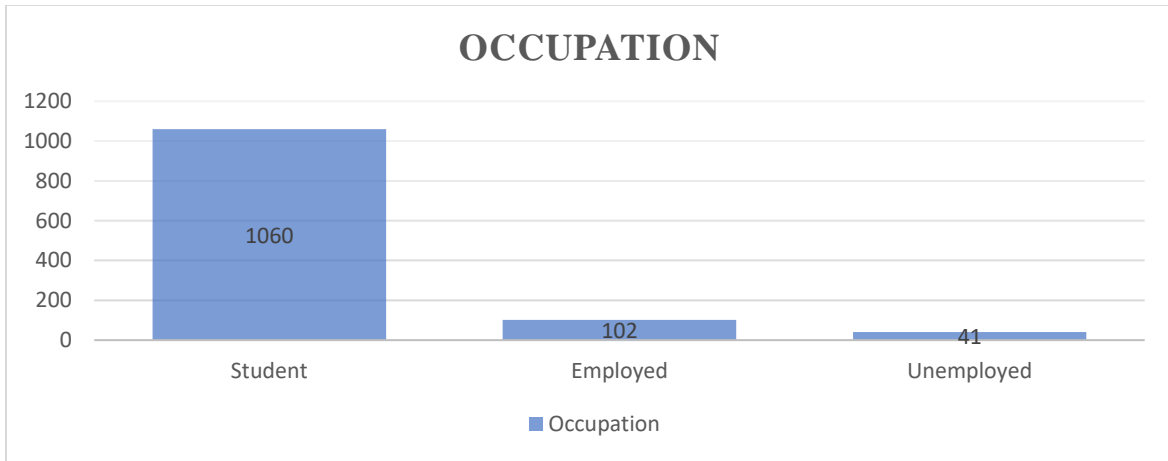


Figure 3.3: Occupation

In this occupation chart from figure 3.3, we can see that majority of the participants are students and their number are 1060. Then the number of employed individuals is 102 and for unemployed, it is 41. So, students play a significant role in this research.

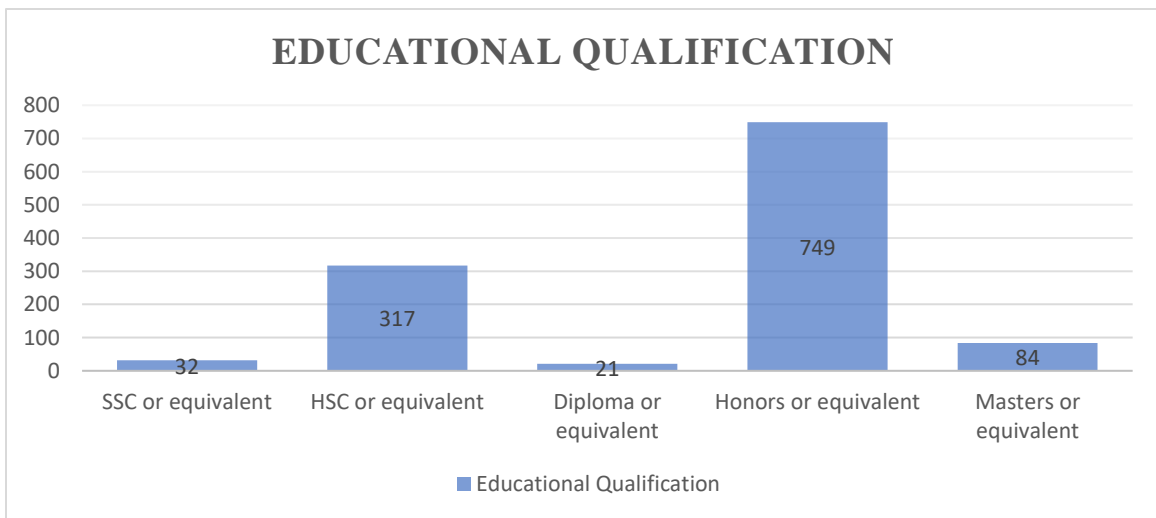


Figure 3.4: Educational Qualification

The educational qualification chart in figure 3.4 shows that 749 participants have an honors degree or equivalent which is the highest among the other degrees. It is followed by people with HSC or equivalent degree where their number is 317. People with Masters or equivalent has also significant participation.

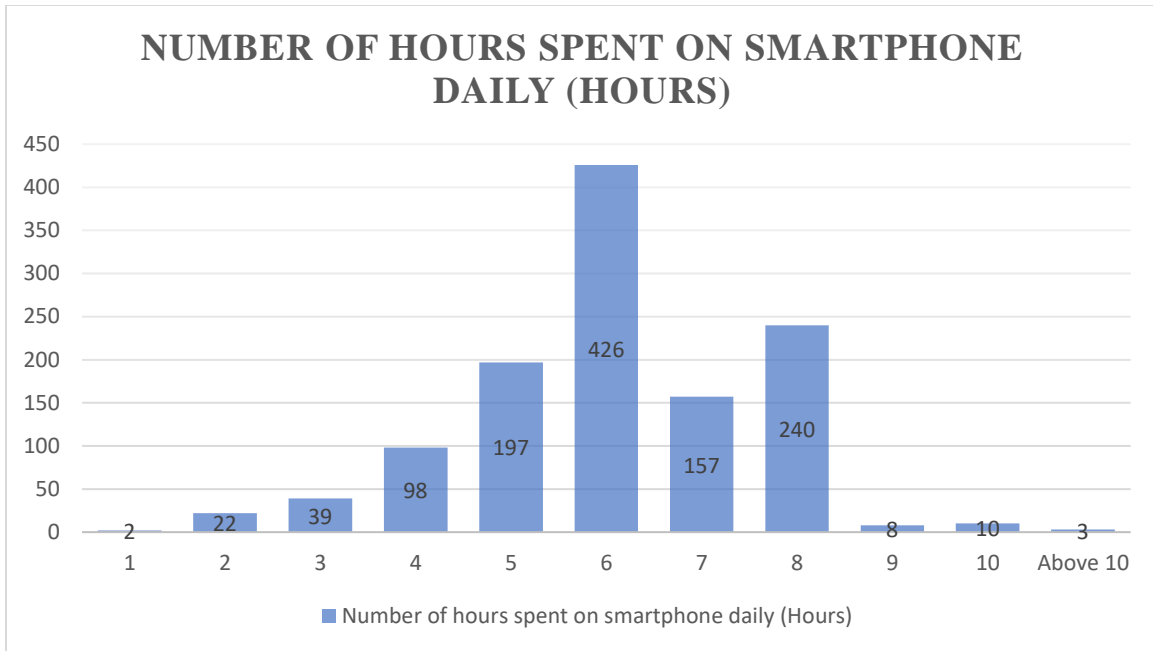


Figure 3.5: Number of Hours spent on smartphone

In the chart from figure 3.5, we can see the time spent daily on smartphone for the participants. Most of the participants, more specifically 426 participants, use their smartphone for 6 hours. Then 240 people use their smartphone for 8 hours and 197 people use for 5 hours. This graph also shows us that majority of the participants use their phone for 4 to 8 hours.

Now, let's look at some other graphs of my dataset:

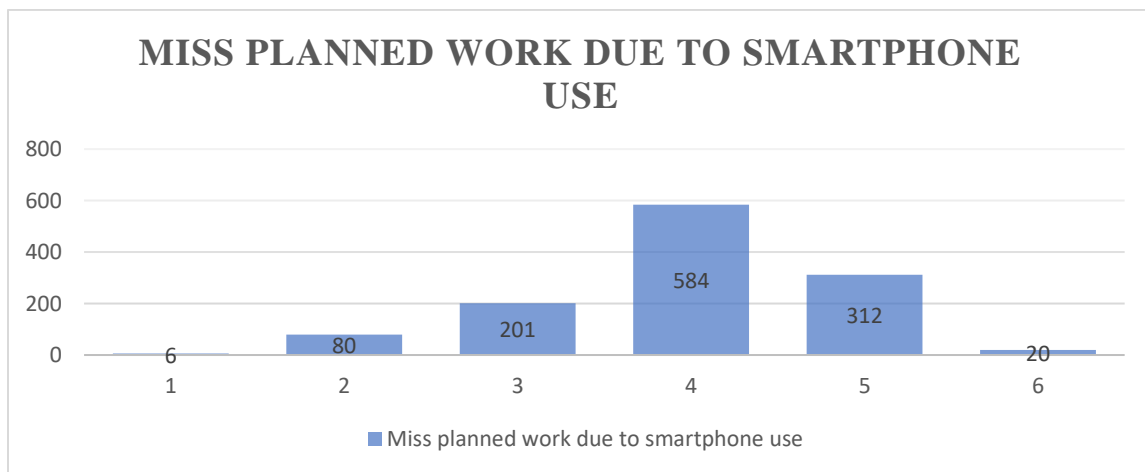


Figure 3.6: Miss Planned Work Due to Smartphone use

Figure 3.6 shows us that, 584 people slightly agreed that they missed their work due to smartphone use and 201 people slightly disagreed in this matter. Also, 20 people strongly agreed with this statement.

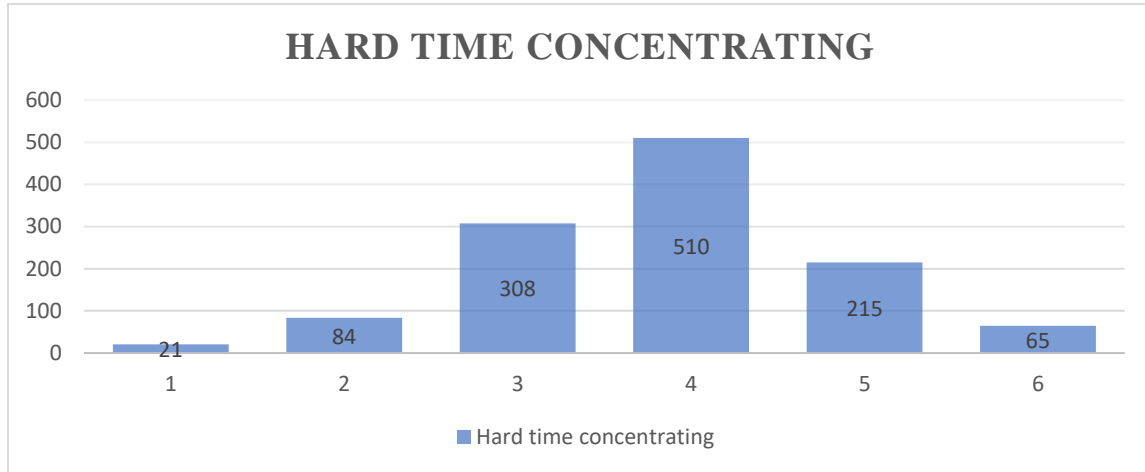


Figure 3.7: Hard time Concentrating

Figure 3.7 shows us that 510 people slightly agreed and 215 people agreed that they had a hard time concentrating because of smartphone overuse. But 308 people slightly disagreed and 84 people disagreed with this statement.

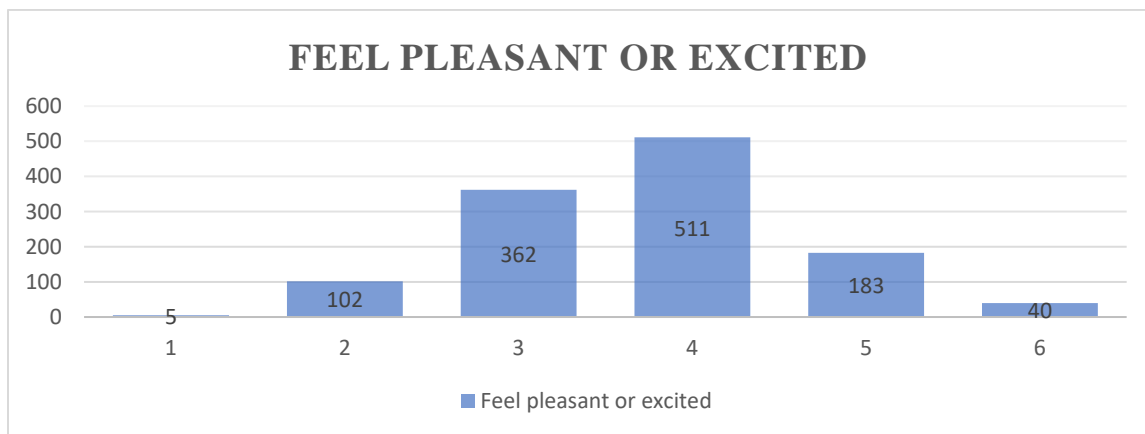


Figure 3.8: Feel pleasant or excited

Figure 3.8 shows us that 511 people slightly agreed that they feel pleasant or excited while using a smartphone but 362 people slightly disagreed in this matter. Also, 40 people strongly agreed with this statement.

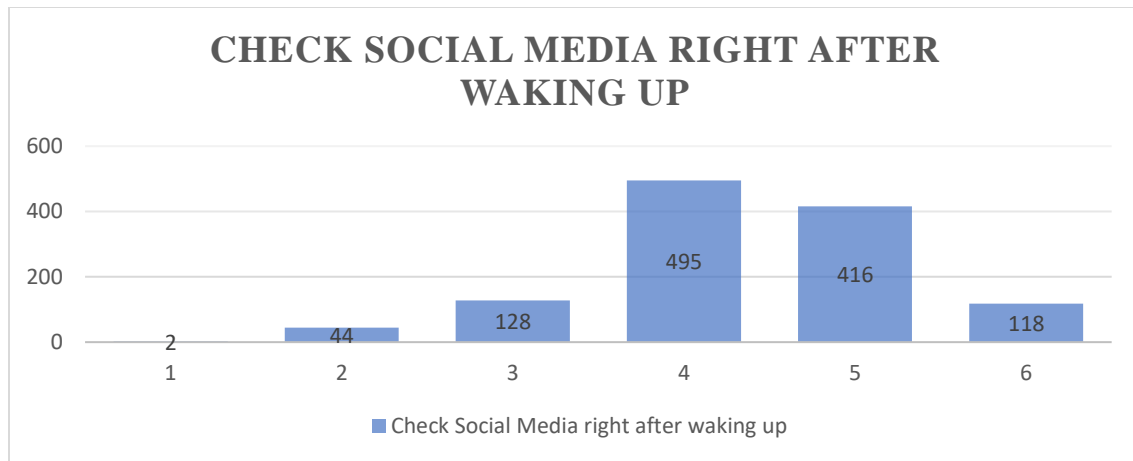


Figure 3.9: Checking social media right after waking up

Figure 3.9 shows that 495 people slightly agreed and 416 people agreed that they had a hard time concentrating because of smartphone overuse. But 128 people slightly disagreed and 44 people disagreed with this statement.

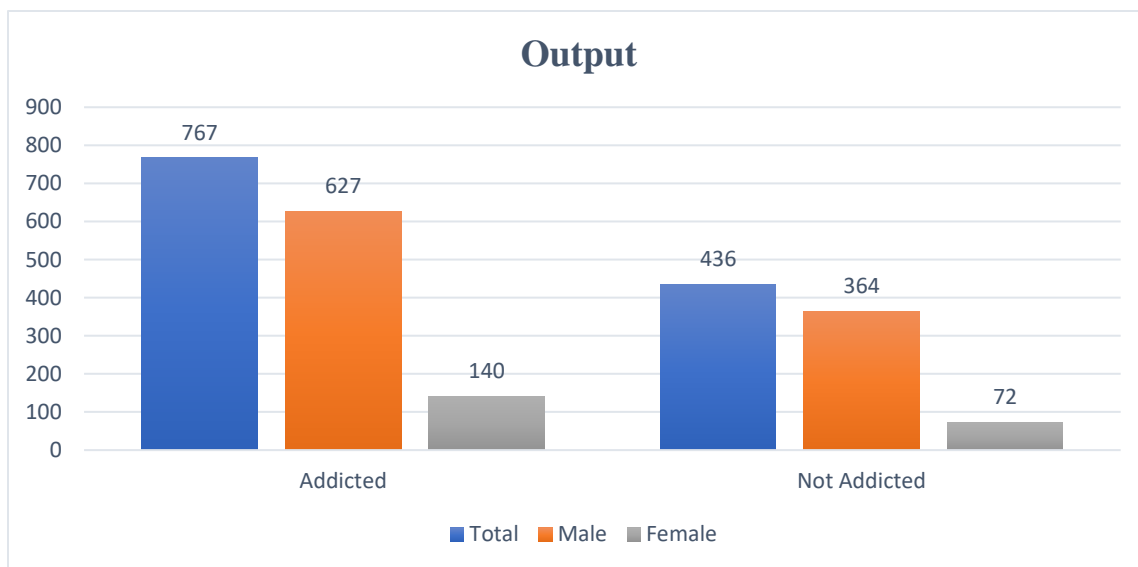


Figure 3.10: Number of addicted and not addicted users

Figure 3.10 shows us the overall output of our dataset. Here, we can see that 762 people are marked as addicted where 627 of them are male and 140 of them are female. On the other hand, 436 people were marked as not addicted where 364 people are male and 72 people are female.

3.4 Proposed Methodology:

This study's methodology began with data collection via an in-depth survey which was designed to collect meaningful information which are related to smartphone addiction prediction. Following that, I performed necessary preprocessing steps on the collected data to ensure its readability which is required for analysis. To standardize the diverse ranges of the used features, I used Standard Scaler technique that enabled consistency and minimized the impact of diverse feature size. Handcrafted labelling was also used to create unique classes or labels based on specific measurers extracted from the features that established the basis for the training of supervised learning models. I divided the dataset into distinct segments, with 80% committed to model training and 20% committed to unbiased model testing and evaluation. I needed to choose this division of data thoroughly to make sure that there is enough sufficiency for training and testing for this dataset which will lead to strong evaluation of the model. In order to build predictive models, I used 6 different machine learning algorithms which are Logistic Regression (LR), AdaBoost (ADB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN) and Decision Tree (DT). Each algorithm was trained on a precise training dataset which is 80% of the original data and learned complicated patterns and relationships that lies between features and their labels. The dataset that was kept for testing purpose was used to carefully evaluate the model performance. I calculated metrics like accuracy, precision, recall and F1-score which gave me the detailed insight of each model's performance and predictive capabilities. After a through comparison of these metrics, I identified the model with best predictive accuracy and reliability which is the best fit for my dataset and this research. This orderly procedure and structured approach guaranteed a detailed examination of different models which helped me chose the most effective and impactful model which is suited for this research's objective. The workflow of my proposed methodology is given below:

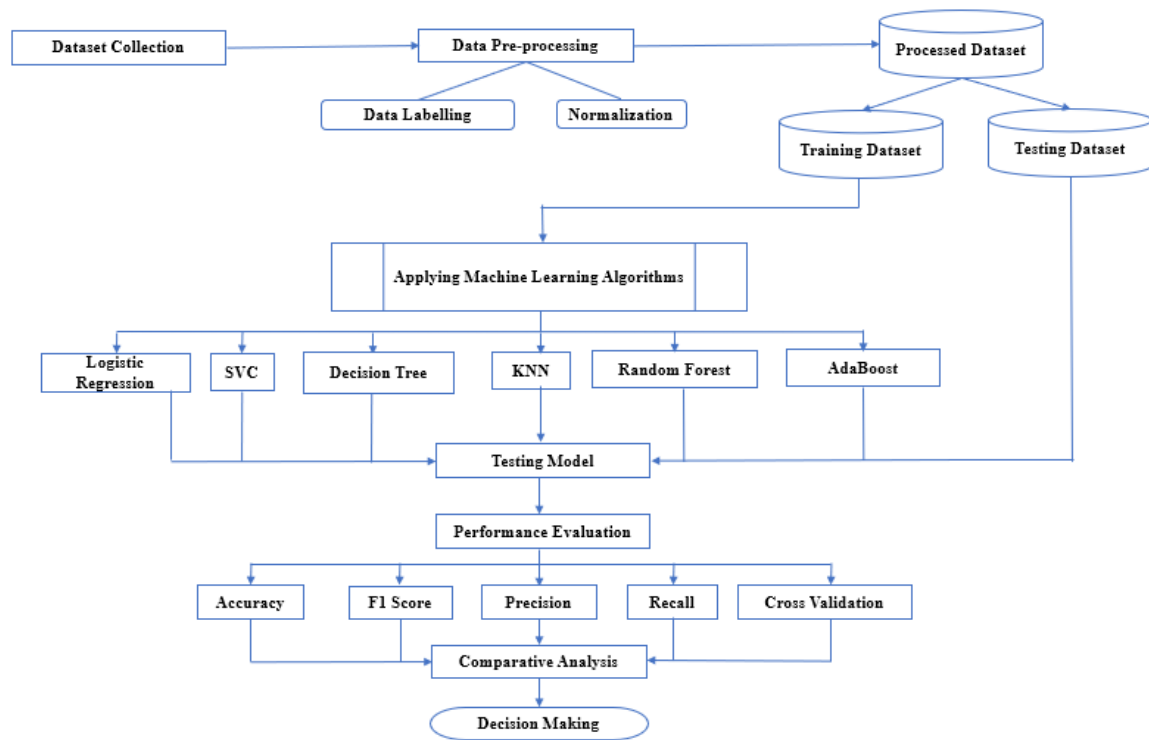


Figure 3.11: Steps of proposed Methodology

3.4.1. Data Pre-Processing:

The data that I collected needed to go through some processes as it has some flaws before using them in the model. As the data was collected through an online form, no data was missing from each response. So, no missing value handling was needed. Also, I only have 6 factors and 20 features to work with. So, I didn't use feature scaling as well. The data that I used was unlabeled. Also, I needed normalization to increase the performance for each model. That's why I used Handcrafted Labeling for data labeling and Standard Scaler for data normalization on the collected dataset.

3.4.1.1. Handcrafted Labeling:

Handcrafted labelling uses human knowledge and expertise into the labelling process, potentially leading to more accurate and specifically relevant data point categorization. As the used dataset is small in size, I used this labeling method in order to achieve more

accuracy and minimize error. I used 2 classes for this dataset which are Addicted and Not Addicted. To categorize each data, I needed a threshold value. For setting the threshold value, I used the mean method. If the average value was equal or greater than the threshold value, I labeled that data as 1 which is Addicted class. And if the average value was less than the threshold value, I labeled the data as 0 which is Not Addicted class. After labeling the data, I proceeded to the next step.

3.4.1.2. Standard Scaler:

After labeling the data, Standard Scaler was applied to the dataset. Standardization is an important step in data preprocessing, particularly when the input variables have different scales and ranges. Because my numerical data ranges from 1 to 6, I scaled it down to have a mean of 0 and a standard deviation of 1. To do so, I used the Standard Scaler, a sci-kit-learn library tool that was adapted to the training data to learn the scaling parameters and then executed on both the training and test datasets. By standardizing the data, I was able to control for scale differences between variables, which could otherwise lead to biased or inaccurate predictive modelling results.

Index	work du	ring assic	rred visic	jack of th	ted while	hone is t	npty with	and not f	hen I am
0	5	2	5	5	4	4	4	3	2
1	3	3	3	3	4	4	4	4	4
2	5	4	6	4	3	5	5	5	3
3	6	5	6	5	5	6	6	6	5
4	5	2	6	6	2	3	1	5	2
5	3	1	2	2	2	2	1	3	1
6	6	5	4	4	5	4	5	4	2
7	5	6	3	5	5	5	4	3	3
8	5	6	4	4	4	4	3	3	4
9	4	4	3	3	5	3	3	4	3
10	5	5	4	4	4	3	3	5	4

Figure 3.12: Before Applying Standard Scaler

Also, I learned that standard scaling has changed the pattern of distribution of my data that distributed the values evenly in a certain range. This standardization process was crucial for the algorithms I planned to use, which are dependent on normally spread data. On the

other hand, figure 3.14 revealed a more evenly distributed values before scaling, which portrayed the original measures accurately but was less valuable for comparative analysis.

	0	1	2	3	4	5
0	0.0843067	1.08003	0.653358	0.727533	0.257841	0.246054
1	0.0843067	1.08003	2.52339	1.69044	2.19165	1.22303
2	0.0843067	-0.738731	0.653358	-0.235378	-0.709063	0.246054
3	2.15365	1.08003	0.653358	0.727533	1.22474	0.246054
4	-0.950366	-0.738731	-1.21668	-0.235378	-1.67597	-0.730925
5	-0.950366	-0.738731	-1.21668	-0.235378	-0.709063	-1.7079
6	0.0843067	0.170649	-0.28166	-0.235378	-0.709063	0.246054
7	1.11898	0.170649	-0.28166	-0.235378	0.257841	0.246054
8	-0.950366	0.170649	-1.21668	-1.19829	0.257841	-0.730925
9	0.0843067	1.08003	-0.28166	-0.235378	1.22474	0.246054
10	1.11898	1.98941	-0.28166	1.69044	1.22474	1.22303

Figure 3.13: After Applying Standard Scaler

Moreover, because of standardization, the impacts of outliers and difference of values of assessments were decreased which helped me see patterns and relationships more clearly.

3.4.2. Selective Algorithms:

In a research-based work where machine learning or data mining techniques are used, selecting the best algorithm is a vital part because the accuracy of the anticipated outcome is heavily dependent on the use of these chosen algorithms. For this research, I used a total of 6 algorithms. They are:

1. Logistic Regression
2. Support Vector Machine
3. Random Forest
4. Decision Tree
5. K Nearest Neighbor
6. Ada Boost Classifier

3.4.2.1. Logistic Regression:

Logistic regression is a classification problem despite its name. It is a statistical technique that is used on binary classification. This algorithm predicts the odds of internal dependency of categorical values. The working process of this algorithm involves replicating the relation between one or more variables which are independent. Also, it predicts the chances of a particular outcome to happen. This algorithm uses the logistic function which is also known as the sigmoid function to contain the predicted output between 0 and 1. This helps in mapping the values that are continuous to the probability score. Moreover, logistic regression uses a linear equation to generate the probability of each binary result. It uses this equation on the input features and then applies a non-linear transformation through this function. The logistic regression equation that is used for a binary classification problem is given below:

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

Here,

- $P(y=1|x)$ is the probability, given the input features x , that the instance to class 1.
- $\beta_0, \beta_1, \beta_2 \dots \beta_n$ are the weights allocated to each feature $x_1, x_2 \dots x_n$
- e is the base of the natural logarithm

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$ is the linear combination that indicates the weighted sum of input features. $1/(1 + e^{-z})$ which is the logistic function converts this sum to a value between 0 and 1. This basically indicates the probability of positive class. In this equation, $\beta_0, \beta_1, \beta_2 \dots \beta_n$ acts as coefficients. We can learn them from the training data using optimization algorithms. The overall purpose of using these optimization algorithms is to figure out the coefficients that work best in corresponding the data and also reduce the margin of error in class prediction.

3.4.2.2. Support Vector Machine:

Support Vector Machine was used in my dataset of smartphone addiction prediction to classify the participants to addicted and not addicted class. It is a widely used supervised learning algorithm that is used for classification tasks. It calculates the best decision threshold or hyperplane that efficiently and successfully separates different classes within the parameter space of the dataset. The reason for using this algorithm in this research is that it always looks for ways to maximize the margin or the distance between the hyperplane and nearest classes which ultimately improves the model's capability to generate and work with new data. In the case of linear data, support vector machine tends to find a linear hyperplane that separates the predictable classes. But if there is nonlinear data, the algorithm will use kernel functions to convert the data into a higher dimensional space which will open the possibility to discover the nonlinear boundaries. The kernel that is used here helps this algorithm to map the input data into high dimension space without calculating the transformation in reality. For the linearly separable data, the hyperplane equation in an SVM is as follows:

$$w \cdot x + b = 0$$

Here,

- The weight vector perpendicular to the hyperplane is denoted by w .
- The input features are represented by x .
- Bias is represented by b .

As support vector machine is highly effective in high dimensional spaces, uses different kernel functions for versatility and works best for both linear and non-linear data, this algorithm works well in classification tasks which is a perfect match for this particular research work.

3.4.2.3. Random Forest:

As an ensemble learning method, Random Forest works best for both regression and classification operations. It basically generates multiple decision trees in the time of training and provides output on the basis of mean prediction of individual trees. The accuracy of algorithms that already exist is unrivalled. It efficiently operates on huge volume of data. It has the capability to manage massive amount of input variables without deleting any variables. As forest development progresses, it produces an intrinsic impartial assessment of generalization error. It comes up with an effective way of computing incomplete data that maintains precision even when a significant portion of the data is missing. It has different methods for balancing errors in unbalanced class data sets. Forests created from other data can be used later. The functionality that are described above can be applied to unlabeled files which results in unsupervised clustering, data views, and outlier detection. It offers an experimental technique for tracking vector interactions. Understanding how they are calculated and generated is quite useful for recognizing and using the various options. Two random forest data artefacts rely on the majority of the options. As the training dataset for the current tree is generated with substitution, roughly 33% of the instances are left out of the dataset. When adding trees to the area, this data is used to perform a neutral assessment of the classification error. It is frequently used to generate predictions of variables that plays a vital role in the dataset. After each tree is generated, all the data scale down the tree, and proximity is computed for each pair of cases. When the same terminal node has been used twice, its proximity increases by 1. At the end, the approximations are scaled by dividing by the number of trees. To replace the lost data and finding outliers, proximities are used. Random forest is a machine learning algorithm used for solving regression and classification problems. These algorithms divide the dataset into many parts and generate a large number of decision trees from it. It makes a decision or predicts an output based on the decision trees which results with the highest probability of the data to appear.

3.4.2.4. Decision Tree:

Decision tree is a supervised learning algorithm. A decision tree algorithm can solve both regression and classification problems, unlike other supervised learning algorithms. The purpose of the Decision Tree is to build a training model, using simple decision rules that are deduced from previous data, which can be used to predict class or value of target variable.

A decision tree has different types of root node, internal and leaf. The decision tree starts at the root node. This root node then branches to internal nodes and leaf. The Leaf node represents the final classification category or the predicted classes. The process of each stem of decision tree is simple and easy, so it is really easy to understand and learn.

In order to build a decision tree, first we need to select a feature that will be the root node. Normally, a single feature can't perfectly guess the last class which we call impurity. To measure this impurity, we use different methods like entropy, Gini and information gain. It also determines how well a feature can separate the classes from a given data. To choose a node, we take the feature with the worst Gini impurity in each level. In order to calculate Gini impurity for a feature, at first, we sort the data in ascending order and then calculate the mean of the neighboring values. After that, we calculated the Gini impurity at each chosen average value by arranging the data points based on whether the feature values are less than or greater than the chosen value and whether that choice correctly classifies the data. The Gini impurity is then calculated using the following equation, where K represents the number of classification categories and P represents the proportion of instances of those categories.

$$Gini\ Impurity = 1 - \sum_{i=1}^K P_i^2$$

After that, we calculate the weighted average of Gini impurity at each leaf. Then the value with the lowest impurity is chosen as the feature. We repeat this process for different features and the value of the selected feature will become the node. This process is repeated at each node and depth level until we classify all of the data. In order to make a prediction for a data point after the tree has been constructed, go down the tree using the conditions

at each node to arrive at the final value or classification. Instead of Gini, the sum of squared residuals or variance is used to measure impurity when using decision trees for regression.

3.4.2.5. K Nearest Neighbor:

K nearest neighbor is also used for both regression and classification task. It is very straightforward but efficient supervised machine learning algorithm. This algorithm works with similarity. It uses a voting method to classify each value. It also uses the average of the K nearest data points where K can be any natural number.

For every new data, this algorithm determines the distance to all other points available in the dataset. To calculate this distance, it mostly uses Euclidian Distance but sometime Manhattan distance is also used. The Euclidian distance formula is as follows:

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

Here, p and q indicate two data points that are present in an n-dimensional space and q_i and p_i indicates the value of each point.

After calculating the value, the algorithm selects the K nearest data points based on their distance. Then it performs the classification task. To do this, it takes a vote from the K neighbors. The class with the major vote will be selected the class of the new query point. For regression, it determines the value by calculating the average of the K neighbors.

K nearest neighbor is an easy-to-understand algorithm and simple to implement. It doesn't consider balanced or unbalanced data distribution so it has the versatility to work with different datasets. Also, because of its working principle, it can adapt to new data which are used for prediction. As a result, it can handle both classification and regression problems. But if the dataset is large, the time and cost to calculate the nearest neighbor will increase which may reduce its performance. Also, it is very sensitive to outliers as it impacts the performance negatively.

3.4.2.6. Ada Boost Classifier:

AdaBoost Classifier is also an ensemble learning algorithm like random forest that is widely used for classification task. The main principle of this algorithm is to combine the weak learners to create a strong learner. These weak learners indicate mainly decision tree or other classifiers. In this algorithm, the weak learners are trained in a sequence where it focuses on adjusting the classifying instances correctly. It particularly focuses on those instances that were previously incorrectly classified. After each iteration, this algorithm adds weight to correctly classified instances that ultimately increase learning process. Then it gives its prediction by combining the weighted voting.

In AdaBoost, weight updates are made by modifying instance weights according to each iteration's classification accuracy. If $D_t(i)$ is the weight of the i^{th} instance at iteration t , α_t is the weight associated with the classifier at iteration t and err_t is the weighted error of the classifier at iteration t , The weight update equation for the $(t+1)^{\text{th}}$ iteration is:

$$D_{t+1}(i) = \frac{D_t(i) * \exp(-\alpha_t * y_i * h_t(x_i))}{Z_t}$$

Here,

- The i^{th} instance's true class label is denoted by y_i .
- $h_t(x_i)$ is what the t^{th} classifier predicts will happen to the i^{th} instance.
- Z_t makes sure that D_{t+1} forms a probability distribution by acting as a normalization factor.

All of the weak learners' predictions combined, weighted, determine AdaBoost's final prediction. For a new instance x , the output $H(x)$ is computed as:

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t \times h_t(x) \right)$$

Here,

- T represents all of the weak learners.
- α_t is the t -th classifier's associated weight
- $h_t(x)$ is the t classifier's prediction for example x .

These formulas embody the fundamental computations and modifications in AdaBoost, facilitating the algorithm's iterative emphasis on misclassified instances and the amalgamation of predictions to generate a robust ensemble classifier.

AdaBoost classifier works best in the process of increasing accuracy. This algorithm can also work with complex datasets by combining the weight of the weak learners which makes them strong learners that increases their influence on the dataset

3.5 Implementation Requirements:

After completing the data collection and data preprocessing, I proceeded to the next part which is implementation. While implementing this data, I had to follow some specific requirements. First of all, the dataset had to be well structured and needed to be adjusted with my system. Then, every algorithm and its associated libraries should be imported into the system to perform the task smoothly. Last but not the least, every testing parameter must be configured to determine the outcome.

The hardware and software used are given below:

- Processor: Intel Core i7 11700
- Ram: 16 GB
- Storage: 512 GB SSD
- Application: Anaconda Navigator with Spyder
- Operating System: Windows 11 Pro

Used Libraries and tools:

- Python 3.11
- Pandas
- Sklearn
- Matplotlib

Chapter 4

EXPERIMENT RESULT AND DISCUSSION

4.1. Experiment Setup:

After completing the data preprocessing part, the most important step of this research approaches which is setting up the experiment. In this part, I used the machine learning algorithms on the preprocessed dataset. To perform this, First I created a simulation-based setup. In this case, I used Spyder. I made sure the latest version of this application was installed. After that, I called the necessary library functions and tools to complete the setup. After setting up the necessary library functions and tools, I loaded the dataset and move towards the result and analysis of the dataset and model performance.

I used the data that I gathered from google form which consists of 20 features and 6 factors. I applied a total of 6 machine learning algorithms for analysis purpose. As they are very well knowing machine learning algorithms, almost all of them performed very well on this dataset.

4.2. Experiment Performance Evaluation Techniques:

When it comes to evaluate the performance of a model, the concepts that are considered fundamental are Confusion Matrix, Accuracy, Recall, Precision and F1 score. The explanation of each term is as follows.

4.2.1. Confusion Matrix:

Confusion matrix is one of the main and fundamental tools for evaluating a model's performance. It is represented in a table that gives us meaningful insight of a model's errors, performance and weaknesses. A confusion matrix has four parts. They are True Positive which is number of correctly predicted positive values, True Negative which is the number of correctly predicted negative values, False Positive which is the number of

incorrectly predicted positive values and lastly False Negative which is the incorrectly predicted negative values. The basic structure of a confusion matrix is given below:

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 4.1: Confusion Matrix

4.2.2. Confusion Matrix Terminology:

Precision: It is the number that measures the ration of number of true positives and the total number of positives that are predicted. The equation goes:

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

Recall: It is also known as Sensitivity. It measures the ratio of true positives and actual positive values.

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

Accuracy: It measures the total number of correctly predicted values and total number of values.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

F1 Score: It measures the harmonic mean of recall and precision.

$$F1\ Score = \frac{2 * Precision * Recall}{Recall + Precision}$$

4.3. Experiment Result and Analysis:

In this section, I will discuss the performance of each algorithm. I evaluated the performance of every algorithm by using Confusion Matrix, ROC Curve and AUC score, Accuracy, Precision, Recall, F1 score and Cross Validation score. The brief explanation of each evaluation method for every algorithm is mentioned in this section. After that, I selected the best algorithm based on these evaluation methods which demonstrates each algorithms performance.

4.3.1. Logistic Regression:

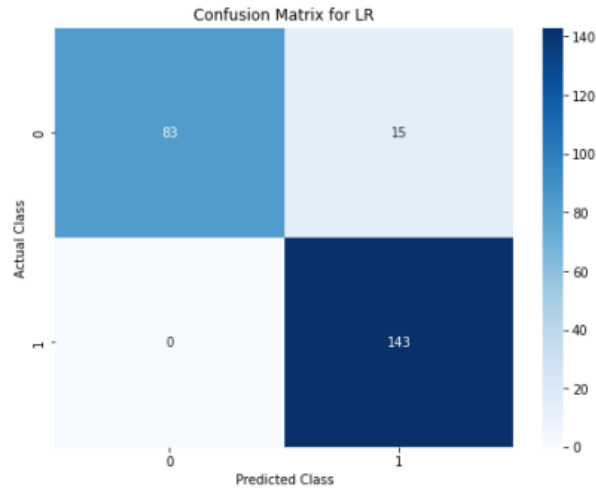


Figure 4.2: Confusion matrix for LR

In figure 4.2, we can see the confusion matrix for logistic regression model. This model correctly predicted 143 positive classes and 83 negative classes. Also, for this model the number of incorrectly classified classes is low as the value of false positive is 15 and value of false negative is 0. With compared to other models, this model's performance is not up to the mark as other models performed better than this model.

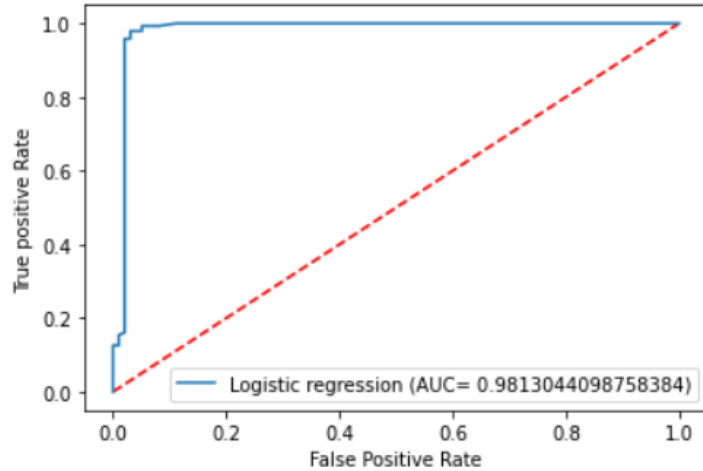


Figure 4.3: ROC Curve for LR

From figure 4.3, we can see the ROC curve for the logistic regression model. The area under the curve or AUC score for this model is 0.981 which is very impressive as the value is closer to 1 which shows good performance. This model's performance is relatively better than other models.

4.3.2. AdaBoost:

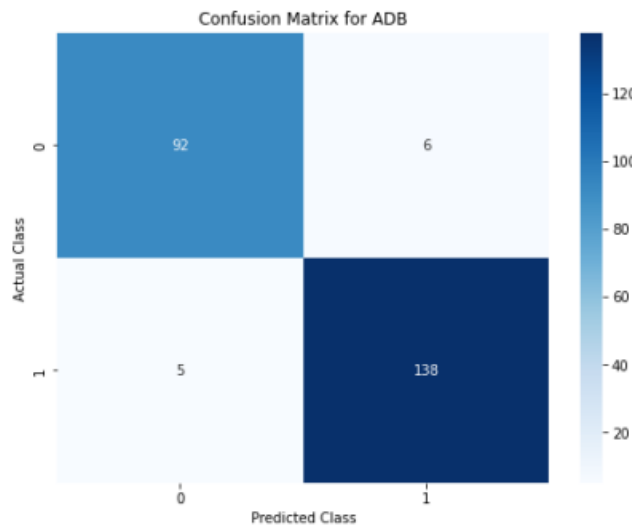


Figure 4.4: Confusion matrix for ADB

In figure 4.4, we can see the confusion matrix for AdaBoost model. This model also performed particularly well as it correctly predicted 138 positive classes and 92 negative

classes. The number of incorrectly predicted class is also low as it only predicted 11 incorrect positive and negative classes which indicates its overall outstanding performance.

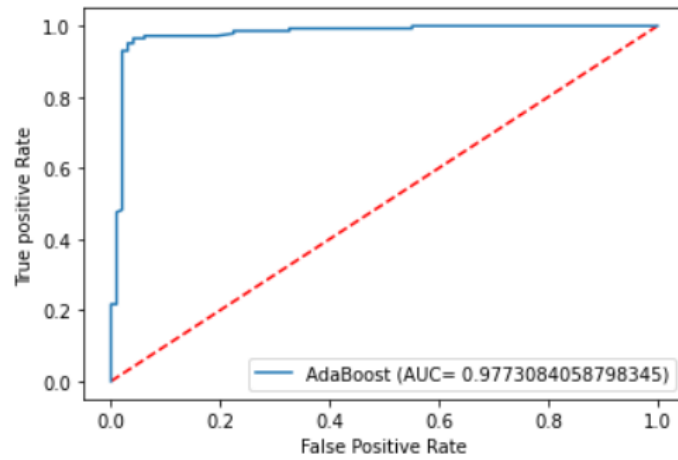


Figure 4.5: ROC curve for ADB

Figure 4.5 shows the ROC curve for AdaBoost model. We can see that the AUC score for this model is 0.977 which is higher than logistic regression which is very impressive. It means it can distinguish between classes very well.

4.3.3. Support Vector Machine:

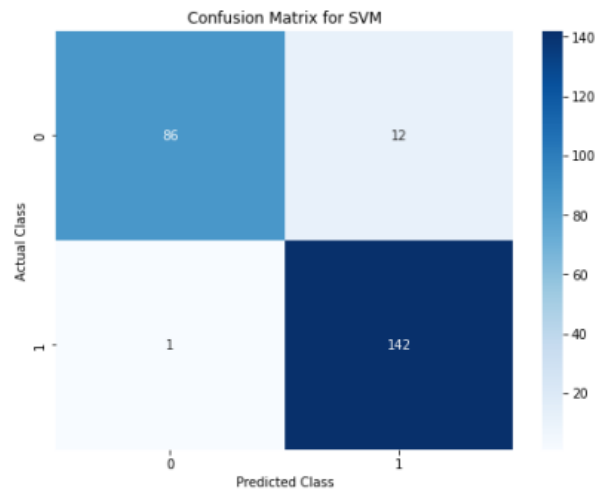


Figure 4.6: Confusion matrix for SVM

The support vectors machine model also shows better performance as the research is performed on binary classification. Figure 4.6 illustrates the confusion matrix for support vectors machine which shows that this model correctly predicted 142 positive and 86 negative classes and incorrectly predicted 12 positive and 1 negative classes.

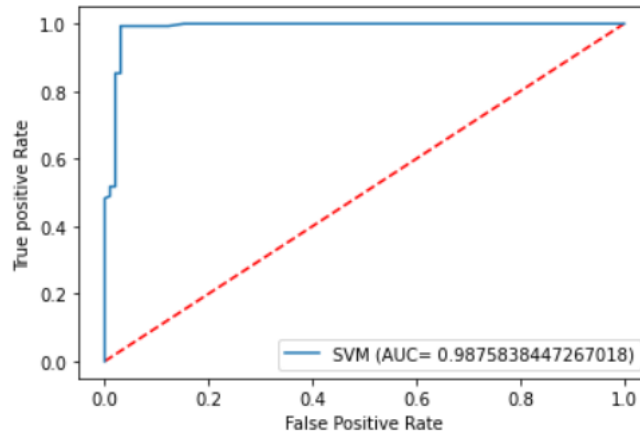


Figure 4.7: ROC curve for SVM

The figure 4.7 of ROC curve for support vector machine model shows better performance than logistic regression as well as AdaBoost model. The AUC score for this model is 0.987. That means this model also performs really good.

4.3.4. Random Forest:

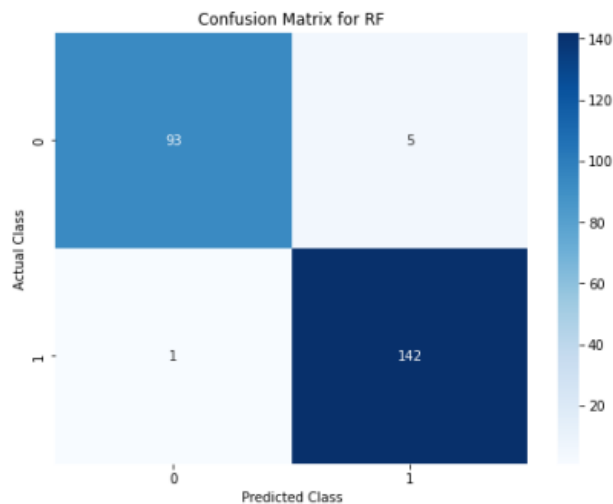


Figure 4.8: Confusion matrix for RF

The random forest model performs the best compared to other models. Figure 4.8 shows that, this model correctly predicted 142 positive classes and 93 negative classes. Also, the number for incorrectly predicted class is 5 for false positive and 1 for false negative.

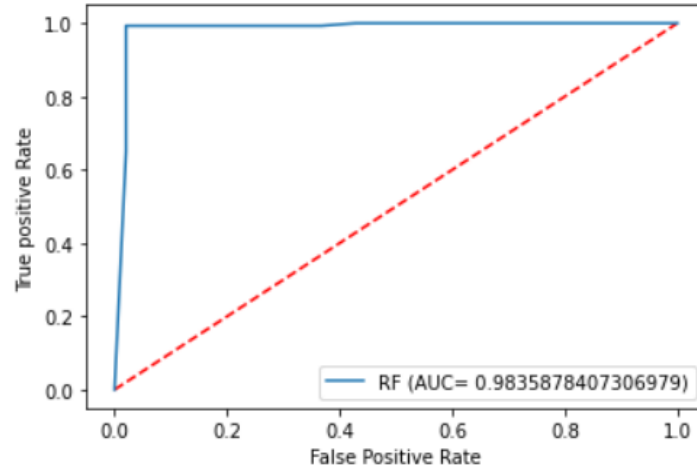


Figure 4.9: ROC curve for RF

The performance of ROC curve for random forest model is also outstanding compared to other models. Figure 4.9 shows that, it achieved 0.983 AUC score. The features of this dataset helped this model to perform this well.

4.3.5. K Nearest Neighbor:

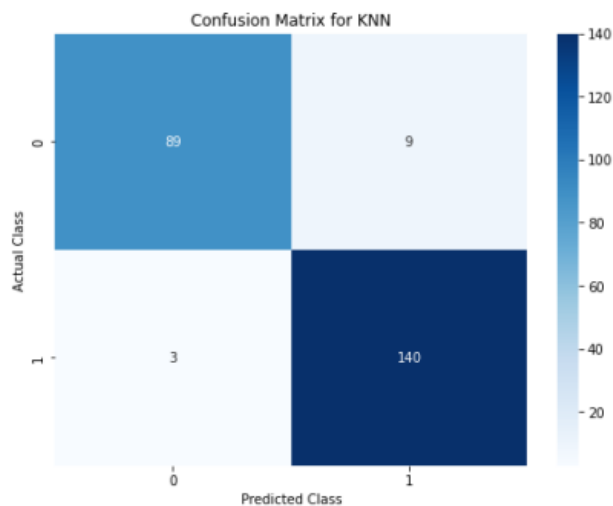


Figure 4.10: Confusion matrix for KNN

From the figure 4.10, we can see that, the k nearest neighbor model can correctly predict 140 positive classes and 89 negative classes from the dataset. Also, for errors, it incorrectly predicted 9 classes as negative which should be positive and 3 classes as positive which should be negative. Overall, the performance of this model shows good promise.

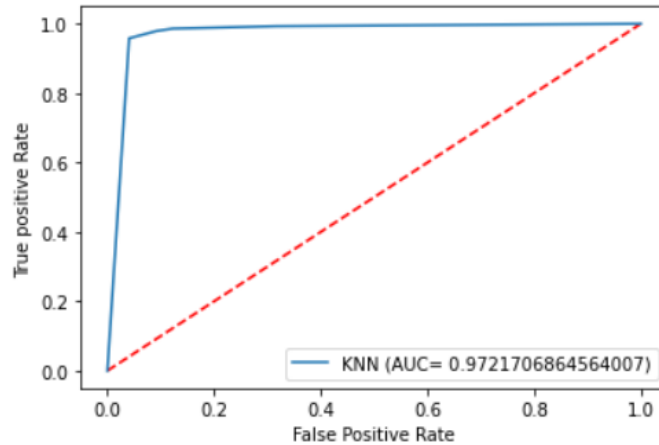


Figure 4.11: ROC curve for KNN

The ROC curve of k nearest neighbor also shows great performance which is indicated in figure 4.11. The AUC score for this model is 0.972 which is not as good as other models but individually it shows outstanding performance.

4.3.6. Decision Tree:

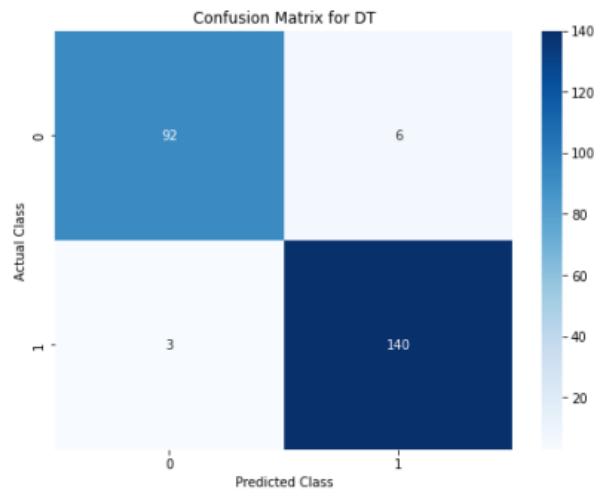


Figure 4.12: Confusion matrix for DT

The figure 4.12 shows the confusion matrix for decision tree model. This model correctly predicted 140 positive class and 92 negative class. It also incorrectly predicted 3 positive class and 6 negative class. This model's performance is also excellent as it didn't predict many incorrect classes.

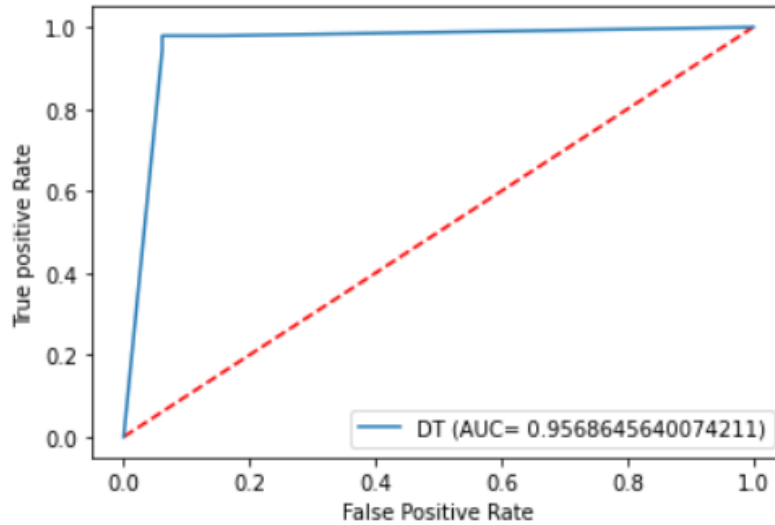


Figure 4.13: ROC curve for DT

The ROC curve of decision tree model which is indicated in figure 4.13 shows that this model has the lowest AUC score among the other models. The AUC score for this model is 0.956. it is not as close to 1 as other model. The bias toward the dominate class might cause this downgrade in performance.

4.3.8. Model Performance Evaluation:

The following table shows the performance evaluation methods which are Accuracy, Precision, Recall, F1 score and Validation Accuracy for each algorithm used in this research:

TABLE 4.1: COMPARISON OF MODEL PERFORMANCE

Algorithm	Accuracy (%)	Validation Accuracy (5-Fold) (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	94.19	97.59	95	92	93
AdaBoost	95.83	95.68	95	95	95
Support Vector Machine	94.60	97.50	96	94	94
Random Forest	97.51	97.67	98	97	97
K Nearest Neighbor	95.02	95.93	95	94	95
Decision Tree	96.26	96.09	96	96	96

From the table 4.1, we can see the performance of each algorithm that I used for this research. To evaluate the performance, I used different evaluation measures like accuracy, cross validation accuracy, precision, recall and f1-score. I used 80% data for training and 20% data for testing for this experiment. If we look at the table, we can see that Random Forest has outperformed other classes. This algorithm achieved the highest accuracy, precision, recall and f1 score which are respectively 97.51%, 98%, 97% and 97%. On the other hand, Logistic Regression has performed the worst among the other algorithms achieving the accuracy of 94.19%, precision of 95%, recall of 92% and f1-score of 93%. It means that for logistic regression, the logistic function is having a hard time capturing the

underlying patterns in the dataset. That's why it performed poorly. If we look at the cross-validation accuracy of the algorithms, we can see that Random Forest has the highest validation accuracy which is 97.67% and AdaBoost has the lowest which is 95.68%. It means that Random Forest will work best with unseen data and it is the most suitable model to work with real-time data among the other algorithms. On the other hand, AdaBoost won't be able to work with real-time data as efficiently as other models as it has the lowest validation accuracy which indicates overfitting.

4.3.9. Best Performing Algorithm:

The best performing algorithm was chosen based on their performances that I evaluated using accuracy, precision, recall, f1-score and cross validation accuracy. From the used algorithms in this research, Random Forest performed the best and was chosen the best algorithm for this research. It achieved top spot by a small margin. Apparently, Random Forest and Decision tree has almost similar accuracy, recall, precision, f1-score and AUC score. But the deciding factor was the cross-validation accuracy. The validation accuracy of decision tree is lower than its initial accuracy. Which means it has slight overfitting problem. This model will perform poorly in terms of unseen data. But the validation accuracy of random forest is greater than the validation accuracy of decision tree. So, this model will perform well in terms of unseen data. Also, graph for the AUC score of Random Forest shown in figure 4.9 shows an almost square shape which proves its overwhelming performance. For these reasons, Random Forest was chosen the best algorithm for this research.

4.4. Discussion:

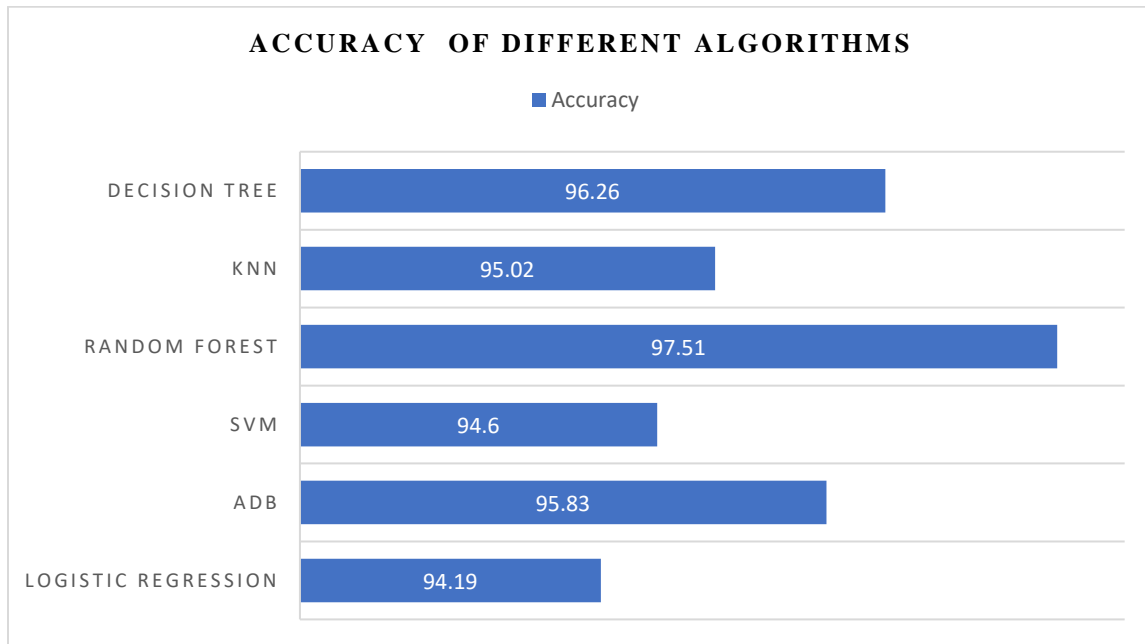


Figure 4.14: Accuracy of different algorithms

In this section, I will discuss the performance of each algorithm that I used in this research. While analyzing the performance of each model, I noticed different implications on my dataset for different algorithms. First of all, Logistic Regression performed poorly in this research achieving 94.19% accuracy, 97.59% validation accuracy, 95% precision, 92% recall and 93% f1 score. It is also the worst performing algorithm of this research. As it is a linear model, it uses a linear function to find patterns which it couldn't do it efficiently for this dataset. That's why Logistic Regression's performance was exceptionally underwhelming. Then the performance score of AdaBoost Classifier is also impressive with 95.83% accuracy, 95.68% validation accuracy, 95% precision, 95% recall and 95% f1-score. As it is an ensemble learning method, it combines multiple weak learners to strong learner. As a result, it performs so well in this dataset. Support Vector Machine also performs quite well as it has 94.6% accuracy, 97.5% validation accuracy, 96% precision, 94% recall and 94% f1-score. This model couldn't perform well in compared to other models because for this dataset, this algorithm is having trouble finding the suitable parameters. That's why this algorithm is not performing as intended. After that, Random Forest achieved 97.51% accuracy, 97.67% validation accuracy, 98% precision, 97% recall

and 97% f1-score. These results suggest that this model performs surprisingly well with this dataset. This also the best performing algorithm in this research as random forest can handle high dimensional data pretty well. The score K Nearest Neighbor achieved is also impressive. It got 95.02% accuracy, 95.93% validation accuracy, 95% precision, 94% recall and 95% f1-score. This score is lower than the other models which means this model is not used to this dataset. Lastly, we have Decision tree where it achieved 96.26% accuracy, 96.09% validation accuracy, 96% precision, 96% recall and 96% f1-score. This is second best performing algorithm in this research. This impressive performance is achieved because this algorithm works particularly well in binary classification data and the dataset used in this research is binary. Also, the medium size of the dataset helped this algorithm to perform well.

Chapter 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1. Impact on Society:

The social impact of smartphone addiction prediction using machine learning while using raw data is significantly noteworthy. This study acts as a source of awareness. This study also illustrates the efficacy of these algorithms in detecting smartphone addiction patterns. It brings up awareness by showcasing the capabilities of machine learning algorithms to detect patterns that are connected to smartphone addiction. Moreover, it initiates a discussion about the safe and responsible use of technology that could ultimately reassess the cultural norms that are interconnected with obsessive smartphone use. It inspires individuals to detect and confront behaviors that are tied to potential compulsive behavior originating from their smartphone usage. If people acknowledge that they are addicted to smartphone, they can adjust their habit before this addiction becomes severe and causes more damage. Furthermore, the study's outcome has the capability to modify the already existing policies and methods that are aimed at reducing smartphone addiction. By prioritizing responsible smartphone use and valuing mental well-being, it clears the way for a culture that embraces a healthier relationship with smartphones. It also focuses on encouraging better habits and attitudes toward tech, creating a more balanced and mindful approach to our digital lives. Finally, this study has the possibility to start an impactful and significant societal change where everyone is more conscious and purposeful about their digital interactions that encourages healthier lifestyles in this digital age.

5.2. Impact on Environment:

Although it may seem indirect, the side-effects of smartphone addiction prediction on the environment are significant when we look at this matter considering environmental responsibility. We know, the environment is greatly impacted when manufacture, use and disposal of smartphone increases. This research can indirectly influence environmental sustainability. One way of doing that is to increasing awareness and promoting more

responsible smartphone use. If the regular use of smartphone decreases through thoughtful smartphone usage, the need for regular smartphone upgrade will go down. As a result, the electronic waste will go down which will directly affect the environment. Also, if people become more conscious about their digital habits, the use energy-intensive server farms and data center will also decrease that will lower overall energy consumption. Moreover, this research can encourage people to innovate eco-friendly device design and material which will also promote the idea of responsible technology use. Initially, this research might have negligible environmental impact, it has the potential to spark a cultural revolution towards more environmentally responsible and friendly technology use that can provide long-term benefits.

5.3. Ethical Aspects:

The model that is used for predicting smartphone addiction does not violate human rights or act in an anti-moral manner. There won't be a privacy issue because the model's data hasn't made publicly available. Rather than undermining an individual's freedom to use or enjoy something, this model contributes significantly to raising awareness. The risk of smartphone addiction prediction model was developed with consideration for all kinds of regulations as well as concerns about confidentiality and privacy. Therefore, the model for predicting smartphone addiction can be managed without any issues using machine learning technology.

5.4. Sustainability Plans:

A few essential steps are part of the sustainability plans for this study. First and foremost, the goal is to educate people on the more energy-efficient and practical ways to use smartphones. Secondly, it includes working with smartphone manufacturers to develop environmentally friendly products. Thirdly, it involves developing more energy-efficient methods for computers to learn. Finally, it requires appealing to policymakers for regulations that encourage phone manufacturers to produce environmentally friendly

products. These efforts try to ensure that smartphone use does not negatively impact the environment and to develop them in ways that are more beneficial to the planet.

Chapter 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1. Summary of the Study:

I looked into the important problem of smartphone addiction in this study, using machine learning to guess how addicts will act. The Smartphone Addiction Scale (SAS) was the basis of my study, and I carefully chose 26 questions that were relevant. Since most of the interviewees were students, the data set is representative of younger people. This set of data was defined by six unique factors that showed how bad the addiction was. These factors included problems in daily life and withdrawal signs. I used different machine learning algorithms to process this data. My goal was to not only figure out the extent of smartphone addiction is, but also to find trends that could help with early intervention plans. My objective was to learn more about smartphone addiction in order to help create tools and methods that would encourage better usage habits and possibly lower the risks of addiction. My study's effects include making people more aware of smartphone addiction and possible treatments, as well as inspiring more responsible smartphone use for the sake of sustainability in society and the environment.

6.2. Conclusion:

In my research, I discovered that machine learning, especially the Random Forest algorithm, is highly effective in predicting smartphone addiction, achieving a remarkable 97.51% accuracy in our study. Although the Logistic Regression algorithm also shows potential, it requires further refinement for better effectiveness. Through this study I highlight the widespread issue of smartphone addiction and showcases the significant role of data-driven methodologies in tackling such contemporary challenges. By leveraging advanced algorithms, we can gain deeper insights and develop more targeted strategies to address the growing concern of smartphone dependency which eventually open up the opportunity for further investigation and deep analytical research in such sensitive issue.

6.3. Implication for Future Study:

My study opens up new areas for research, especially when it comes to making prediction models better for people from different backgrounds. It shows that data-driven methods can help us understand behavioral problems, which suggests they could be used in more situations. Future researchers can use these results from my work to look into more complex aspects of digital addiction, make better prediction models, and test different ways to help people who are addicted. Another thing that can be looked into in the future is how cultural, social, and work factors affect smartphone addiction. Also, there is room to look into intervention methods that can work well with predictive models to lower the number of people who become addicted. Future study can improve the accuracy and usefulness of these predictive models by exploring different machine learning algorithms and adding more data to the dataset. Lastly, this work makes people want to use similar methods to study other psychological or behavioral instances. This could lead to personalized and effective solutions for digital well-being.

References

- [1]. Wilmer HH, Sherman LE, Chein JM. "Smartphones and cognition: A review of research exploring the links between mobile technology habits and cognitive functioning." *Frontiers in psychology*, pp 8:605. April 25, 2017.
- [2]. Hale L, Guan S. "Screen time and sleep among school-aged children and adolescents: a systematic literature review." *Sleep medicine reviews*, pp. 50-58, June 1,2015.
- [3]. Mazumdar A, Karak G, Sharma S. "MACHINE LEARNING MODEL FOR PREDICTION OF SMARTPHONE ADDICTION." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, pp 545-550, 2020.
- [4]. Aljomaa SS, Qudah MF, Albursan IS, Bakhiet SF, Abduljabbar AS. "Smartphone addiction among university students in the light of some variables." *Computers in Human Behavior*. Pp 155-164, August 1. 2016.
- [5]. Arora, Anshika, et al. "Intelligent Model for Smartphone Addiction Assessment in University Students using Android Application and Smartphone Addiction Scale." *International Journal of Education and Management Engineering*. P 29, February 1, 2023.
- [6]. Shin C, Dey AK. "Automatically detecting problematic use of smartphones. InProceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing." pp. 335-344, September 8, 2013.
- [7]. Lee, Juyeong, and Woosung Kim. "Prediction of problematic smartphone use: A machine learning approach." *International journal of environmental research and public health* 18.12 (2021): 6458.
- [8]. TASPINAR, Gulsen. "KOKLU N., TASPINAR G., SULAK SA, (2021). Predicting Smartphone Addiction with Support Vector Machine, *Akademik Arastirmalar* 2021, (2), pp. 56-64, Cizgi Kitabevi, Konya, TURKEY. Nigmet KOKLU Konya Technical University."
- [9]. Giraldo-Jiménez, Claudia Fernanda, et al. "Smartphone's dependency risk analysis using machine-learning predictive models." *Scientific Reports* 12.1 (2022): 22649.
- [10]. Bisen, Shilpa, and Yogesh Deshpande. "An analytical study of smartphone addiction among engineering students: a gender difference." *The International Journal of Indian Psychology* 4.1 (2016): 70-83.
- [11]. Achal, Fairuz Tanzim, Mosammat Suraiya Ahmmed, and Tanjim Taharat Aurpa. "Severity Detection of Problematic Smartphone Usage (PSU) and its Effect on Human Lifestyle using Machine Learning." *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*. IEEE, 2023.

- [12]. Tateno, M., Kim, D. J., Teo, A. R., Skokauskas, N., Guerrero, A. P., & Kato, T. A. (2019). Smartphone addiction in Japanese college students: usefulness of the Japanese version of the smartphone addiction scale as a screening tool for a new form of internet addiction. *Psychiatry investigation*, 16(2), 115.
- [13]. Uz Zaman, N., Akther, A., Tabassum, N., Samrat, M. K. K., & Khan, S. M. (2023). An approach to detect smartphone addiction through activity recognition and app usage behaviour (Doctoral dissertation, Brac University).
- [14]. Cha, S. S., & Seo, B. K. (2018). Smartphone use and smartphone addiction in middle school students in Korea: Prevalence, social networking service, and game use. *Health psychology open*, 5(1), 2055102918755046.
- [15]. Aggarwal, S., Gupta, S., Satia, S. P. S., Saluja, S., & Gambhir, V. (2022). Pilot Study to Predict Smartphone Addiction Through Usage Pattern of Installed Android Applications and to Derive Correlations Between Addiction and Phone Usage Behavior. *Addicta: The Turkish Journal on Addictions*, 9(1).
- [16]. Islam, M. Z., Jannat, Z., Habib, M. T., Rahman, M. S., & Islam, G. Z. (2022, May). Detection of Facebook addiction using machine learning. In *International Conference on Image Processing and Capsule Networks* (pp. 625-638). Cham: Springer International Publishing.
- [17]. Eichenberg, C., Schott, M., & Schroiff, A. (2021). Problematic smartphone use—comparison of students with and without problematic smartphone use in light of personality. *Frontiers in Psychiatry*, 11, 599241.
- [18]. Kwon, M., Kim, D. J., Cho, H., & Yang, S. (2013). The smartphone addiction scale: development and validation of a short version for adolescents. *PloS one*, 8(12), e83558.
- [19]. Jun, W. (2015, December). An analysis study on correlation of internet addiction and school age. In *2015 2nd International Conference on Information Science and Security (ICISS)* (pp. 1-3). IEEE.
- [20]. Lee, J. K., Kang, H. W., & Kang, H. B. (2015). Smartphone addiction detection-based emotion detection result using random Forest. *Journal of IKEEE*, 19(2), 237-243.
- [21]. Liu, H., Zhou, Z., Huang, L., Zhu, E., Yu, L., & Zhang, M. (2022). Prevalence of smartphone addiction and its effects on subhealth and insomnia: a cross-sectional study among medical students. *BMC psychiatry*, 22(1), 1-7.
- [22]. Pi, S. Y. (2013). Self-diagnostic system for smartphone addiction using multiclass SVM. *Journal of the Korean Data and Information Science Society*, 24(1), 13-22.
- [23]. Opoku Asare, K., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., & Ferreira, D. (2021). Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. *JMIR mHealth and uHealth*, 9(7), e26540.

- [24]. Lin, Y. H., Lin, P. H., Chiang, C. L., Lee, Y. H., Yang, C. C., Kuo, T. B., & Lin, S. H. (2017). Incorporation of mobile application (app) measures into the diagnosis of smartphone addiction. *The Journal of clinical psychiatry*, 78(7), 4399.
- [25]. Akhtar, F., Patel, P. K., Heyat, M. B. B., Yousaf, S., Baig, A. A., Mohona, R. A., ... & Wu, K. (2023). Smartphone addiction among students and its harmful effects on mental health, oxidative stress, and neurodegeneration towards future modulation of anti-addiction therapies: a comprehensive survey based on slr, Research questions, and network visualization techniques. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, 22(7), 1070-1089.
- [26]. Lei, L. Y. C., Ismail, M. A. A., Mohammad, J. A. M., & Yusoff, M. S. B. (2020). The relationship of smartphone addiction with psychological distress and neuroticism among university medical students. *BMC psychology*, 8, 1-9.
- [27]. Li, L., Wang, L., & Wang, X. (2022). Effect of smartphone use before bedtime on smartphone addiction behaviors among Chinese college students. *Frontiers in Psychology*, 13, 1023245.
- [28]. Chiang, H. S., Dong, Z. Y., Chen, M. Y., & Chen, A. P. (2019). Exploring the impact of smartphone addiction in prospective memory. *Journal of Advances in Information Technology Vol*, 10(1).
- [29]. Osorio-Molina C, Martos-Cabrera MB, Membrive-Jiménez MJ, Vargas-Roman K, Suleiman-Martos N, Ortega-Campos E, Gómez-Urquiza JL. "Smartphone addiction, risk factors and its adverse effects in nursing students: A systematic review and meta-analysis." *Nurse Education Today*., p 104741. March 1, 2021.
- [30]. Abu-Taieh EM, AlHadid I, Kaabneh K, Alkhaldeh RS, Kwaldeh S, Masa'deh RE, Alrowwad AA. "Predictors of Smartphone Addiction and Social Isolation among Jordanian Children and Adolescents Using SEM and ML." *Big Data and Cognitive Computing*. P 92, September 2, 2022.

PLAGIARISM REPORT

ORIGINALITY REPORT

17% SIMILARITY INDEX	13% INTERNET SOURCES	7% PUBLICATIONS	7% STUDENT PAPERS
--------------------------------	--------------------------------	---------------------------	-----------------------------

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to University College London Student Paper	1%
4	ebin.pub Internet Source	1%
5	www.mdpi.com Internet Source	1%
6	www.researchgate.net Internet Source	<1%
7	academic-accelerator.com Internet Source	<1%
8	"Emerging Technologies in Data Mining and Information Security", Springer Science and Business Media LLC, 2019 Publication	<1%