

Machine Learning Applications in Early Screening of Depression and Anxiety using RCADS-47



By

Saleha Noor

Registration No: 00000402087

Department of Sciences

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

2024

Machine Learning Applications in Early Screening of Depression and Anxiety using RCADS-47



By

Saleha Noor

Registration No: 00000402087

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Masters of Science in

Bioinformatics

Supervisor: Dr. Zamir Hussain

School of Interdisciplinary Engineering & Sciences (SINES)


National University of Sciences & Technology (NUST)

Islamabad, Pakistan


August 2024

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by **Ms. Saleha Noor** Registration No. **00000402087** of **SINES** has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.


Signature with stamp: 
Name of Supervisor: Dr. Zamir Hussain
Date: Aug 16, 2024

Associate Professor
SINES - NUST, Sector H-12
Islamabad

Signature of HoD with stamp: 
Date: 19-8-2024

Dr. Fozia Malik
HoD Sciences
Professor
SINES NUST Sector H 12
Islamabad

Countersign by

Signature (Dean/Principal): 
Date: 19/08/2024 **Principal SINES**

AUTHOR'S DECLARATION

I, Saleha Noor, hereby state that my MS thesis titled “Machine Learning Applications in Early Screening of Depression and Anxiety using RCADS-47” is my own work and has not been submitted previously by me for taking any degree from the National University of Sciences and Technology, Islamabad, or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Saleha Noor

Date: 5th August 2024

Certificate for Plagiarism


It is certified that the MS Thesis Titled Machine Learning Applications in Early Screening of Depression and Anxiety using RCADS-47 by Saleha Noor has been examined by us. We undertake the following:

- a. Thesis has significant new work/knowledge as compared to already published or is under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph, or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.
- b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results, or words of others have been presented as the Author's own work.
- c. There is no fabrication of data or results which have been compiled/analyzed.
- d. There is no falsification by manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
- e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name of Supervisor:

Dr. Zamir Hussain

**Signature & Stamp of
Supervisor:**


Associate Professor
SINES - NUST, Sector H-12
Islamabad

DEDICATION

This thesis is dedicated to my beloved family, who sacrifice and contribute in many different ways, and my dearest friends Humaira Batool and Mariam Bint Imran.

ACKNOWLEDGEMENTS

My profound gratitude goes to Almighty Allah for the divine blessings that sustained me throughout the demanding yet fulfilling journey of this research. He handled everything that would have prevented me from continuing, gave me strength through my most trying times, and enabled me to overcome challenges and achieve my goals, leading to the successful completion of this project.

I extend my deepest gratitude to my supervisor, Dr. Zamir Hussain, for his support, guidance, and knowledge. His constant motivation, enthusiasm, and patience were indispensable throughout the entire process. I appreciate your assistance and advice and will always be grateful to you.

I am also grateful to the School of Interdisciplinary Engineering and Sciences (SINES), NUST, for providing the necessary research infrastructure and enabling environment for this project. I deeply appreciate the members of my Guidance and Examination Committee (GEC), Dr. Zartasha Mustansar, Dr. Rehan Zafar Paracha, and Dr. Qurrat Ulain Hamdan (External), whose expert guidance, critiques, and support were invaluable throughout the research process.

A special thanks to Dr. Qurrat Ulain Hamdan from the Psychiatry Department of Benazir Bhutto Hospital, Rawalpindi, for graciously providing the data that made this research possible.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	VIII
TABLE OF CONTENTS	IX
LIST OF TABLES	XI
LIST OF FIGURES	XII
LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS	XIII
ABSTRACT	XIV
CHAPTER 01: INTRODUCTION	1
1.1 Mental Health Challenges in Pakistan	1
1.2 Depression and Anxiety	4
1.3 Revised Child Anxiety and Depression Scale (RCADS)	5
1.4 Problem Statement	6
1.5 Objectives	7
1.6 Relevance to National Needs	8
1.7 Thesis Structure	8
CHAPTER 02: LITERATURE REVIEW	9
2.1 Evaluation of Psychometric Properties of RCADS	9
2.1.1 Australia	9
2.1.2 Turkey	10
2.1.3 Hawaii	10
2.1.4 Denmark	11
2.1.5 Netherlands	12
2.1.6 United States of America	13
2.1.7 Ireland	14
2.1.8 United Kingdom	15
2.1.9 Spain	15
2.1.6 Cross-national	16
2.2 Machine Learning for Depression and Anxiety	17
2.3 Study Rationale	20
CHAPTER 03: METHODOLOGY	21
3.1 Data Collection	22
3.2 Data Preprocessing	22
3.3 Data Analysis	23
3.3.1 Descriptive Analysis	23
3.3.2 Internal Consistency and Reliability Analysis	23
3.4 Feature Selection	24

3.4.1 Chi-square Test of Independence	24
3.4.2 Spearman Correlation	24
3.4.3 Recursive Feature Elimination (RFE)	25
3.5 Data Augmentation	26
3.5.1 Multinomial Probability Distribution	26
3.5.2 Data Generation	27
3.6 Model Development	27
3.5.1 Decision Tree	27
3.5.2 Random Forest Tree	27
3.5.3 Support Vector Machine	28
3.5.4 Logistic Regression	28
3.5.5 Naive Bayes	28
3.5.6 K-Nearest Neighbour	29
3.7 Model Evaluation	29
3.6.1 Confusion Matrix	29
3.6.2 Accuracy	29
3.6.3 Precision	30
3.6.4 Recall	30
3.6.6 F1 Score	30
CHAPTER 04: RESULTS AND DISCUSSION	31
4.1 Descriptive Analysis	31
4.2 Internal Consistency and Reliability Analysis	35
4.3 Feature Selection	37
4.4 Data Augmentation	45
4.5 Model Development and Evaluation	47
CHAPTER 05: CONCLUSIONS AND FUTURE RECOMMENDATIONS	60
5.1 Key Findings	60
5.2 Limitations	60
REFERENCES	62
APPENDIX A: THE REVISED CHILD ANXIETY AND DEPRESSION SCALE (RCADS-47)	73
APPENDIX B: PYTHON CODE FOR RF-RFE	75
APPENDIX C: R CODE FOR CHI-SQUARE TEST FOR MULTINOMIAL DISTRIBUTION	78
APPENDIX D: R CODE FOR DATA AUGMENTATION	80
APPENDIX E: PYTHON CODE FOR ML MODELS	82

LIST OF TABLES

	Page No.
Table 1.1: Common Mental Health Assessment Tools	2
Table 1.2: RCADS-47 Subscales and their Corresponding Items	6
Table 4.1: Distribution of data (gender)	33
Table 4.2: Distribution of data (grade)	33
Table 4.3: Prevalence of the subscale disorders in the data.....	33
Table 4.4: Tests of normality of RCADS-47 subscales.....	34
Table 4.5: Skewness and Kurtosis of RCADS-47 subscales	35
Table 4.6: Internal consistency coefficient Cronbach's alpha for each subscale	36
Table 4.7: Results of the Chi-square test of independence of the 47 features	38
Table 4.8: Results of Spearman correlation of the 47 features with the target.....	40
Table 4.9: Selected features from the three feature selection methods.....	43
Table 4.10: Results of the Chi-square test for multinomial distribution.....	46
Table 4.11: Performance evaluation of the six ML models on the original dataset	48
Table 4.12: Performance evaluation of the six ML models on the 1:4 hybrid and feature-selected dataset.....	49
Table 4.13: Performance evaluation of the six ML models on the 1:8 hybrid and feature-selected dataset.....	51
Table 4.14: Performance evaluation of the six ML models on the 1:12 hybrid and feature-selected dataset.....	52
Table 4.15: Performance evaluation of the six ML models on the 1:16 hybrid and feature-selected dataset.....	54
Table 4.16: Performance evaluation of the six ML models on the 1:20 hybrid and feature-selected dataset.....	55
Table 4.17: Summary of performance evaluations of the six ML models on five hybrid datasets	57

LIST OF FIGURES

	Page No.
Figure 3.1: Workflow of the methodology	21
Figure 4.1: Bar chart of distribution (gender).....	21
Figure 4.2: Bar chart of distribution (grade).....	32
Figure 4.3: Heat map of inter-item correlation matrix.....	37
Figure 4.4: Bar chart of important features selected by RF-RFE	42
Figure 4.5: Confusion matrices of the six ML models on the original dataset.....	49
Figure 4.6: Confusion matrices of the six ML models on the 1:4 hybrid dataset.....	50
Figure 4.7: Confusion matrices of the six ML models on the 1:8 hybrid dataset.....	52
Figure 4.8: Confusion matrices of the six ML models on the 1:12 hybrid dataset.....	53
Figure 4.9: Confusion matrices of the six ML models on the 1:16 hybrid dataset.....	55
Figure 4.10: Confusion matrices of the six ML models on the 1:20 hybrid dataset.....	56
Figure 4.11: F1 score comparison of the six ML models across the five datasets	58

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

ML	Machine Learning
AI	Artificial Intelligence
RCADS	Revised Child Anxiety and Depression Scale
LMICs	Low- and Middle-Income Countries
MDD	Major Depressive Disorder
GAD	Generalized Anxiety Disorder
OCD	Obsessive-Compulsive Disorder
SAD	Separation Anxiety Disorder
PD	Panic Disorder
SP	Social Phobia
SE	Standard Error
R	Pearson Correlation Coefficient
χ^2	Chi-square Statistic
C	Contingency Coefficient
df	Degree of Freedom
RFE	Recursive Feature Elimination
RF	Random Forest
DT	Decision Tree
SVM	Support Vector Machine
LR	Logistic Regression
NB	Naive Bayes
KNN	K-Nearest Neighbor
N	Normal class
B	Borderline class
C	Clinical class

ABSTRACT

Depression and anxiety are prevalent among 10-20% of children and adolescents globally, with an estimated 15 million people affected in Pakistan. Despite this growing figure, the general Pakistani population lacks awareness regarding mental disorders due to limited mental healthcare resources and negative perception of mental health. This study aims to utilize machine learning with RCADS to maximize the use of current healthcare resources and facilitate depression and anxiety screening. Three feature selection methods i.e., the Chi-square test of independence, Spearman correlation, and Recursive Feature Elimination revealed a weak correlation with the evaluation of depression and anxiety in the study population. Data augmentation was done using the multinomial probability distribution of the existing data to generate hybrid-synthetic correlated discrete multinomial variates of each item of RCADS-47, to address the limitation of a small sample size. Six commonly employed ML algorithms—Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, Naive Bayes, and K-Nearest Neighbor—were trained on the hybrid data to develop the predictive models. The Naive Bayes algorithm yielded the best overall results with up to 75% accuracy and a 0.75 F1 score. The findings suggest that the Naive Bayes algorithm using 46 features suits the data well and has the potential to be used as a data-driven decision support system for the concerned professionals and improve the usual way of screening anxiety and depression in children and adolescents.

Keywords: Revised Child Anxiety and Depression Scale (RCADS), Machine Learning Algorithms, Depression, Anxiety, Data Augmentation

CHAPTER 01: INTRODUCTION

Adolescence is a chaotic transitional stage marked by significant physiological, psychological, and emotional changes. This maelstrom of change and adaptation makes young people more susceptible to mental health illnesses such as anxiety, mood disorders, eating disorders, and personality disorders [1]. The most prevalent forms of mental health problems in children and adolescents include psychological distress such as depression and anxiety [2], [3] and it has a severe effect on their lives. The rate ranges from 11% to 25% and 3% to 8% for anxiety disorders and depressive disorders respectively [4], [5]. It is predicted that half of all mental problems begin to develop by the age of 14, and 75% by the age of 18 [6], [7]. Untreated anxiety and depression may have negative effects and cause other issues later in life, such as substance misuse or dependency, suicidal thoughts, poor academic performance, and unemployment [8], [9], [10], [11]. The cause for the observed rise in psychological illness among young individuals is unknown. To address this rise, parents, schools, medical professionals, and governments must work together to develop supportive settings and offer the tools and solutions that are needed.

1.1 Mental Health Challenges in Pakistan

Even though the mental health of children is becoming a global priority, there is limited research on the subject from low- and middle-income nations like Pakistan [12]. In Pakistan, around 15 million people struggle with mental health issues. But for a country with a population of 241.49 million [13], there are relatively few government mental health facilities and just 400 certified psychiatrists majority of whom are located in urban areas [14]. Even though about half of Pakistan's population is under the age of 18, as far as we know, no empirical statistics for children and adolescents have been recorded on a national level [15]. Nonetheless, studies and surveys with small sample numbers show that there is a burden associated with early onset mental health problems. A survey conducted in Rawalpindi on 1,124 youth revealed that 17.2% and 21.4% of them were estimated to be suffering from anxiety and depression, respectively [16]. A study conducted in Karachi on 1,470 individuals between the ages of 11 and 17 found that around 20% of the participants

had serious emotional and behavioral issues. Similarly, a survey conducted on 640 teenagers estimated that 34% of the participants had atypical social and emotional behavior [17], [18]. Regardless of the given estimates, resources for addressing mental health disorders are insufficient for their severity. According to the World Health Organization (WHO) data, as of 2019, the suicide mortality rate (per 100,000 population) in Pakistan is 8.9 [19]. When it comes to mental health concerns, the majority of the community appears to be in denial. Individuals are reluctant to disclose that they suffer from mental health issues because these are taboo subjects that are hardly discussed [20]. Mental health is just as important as physical health, yet it is often neglected due to cultural stigma, a lack of awareness, and misleading spiritual beliefs. The general public is unaware of mental illnesses and the small percentage that is informed is unaware of the therapies available for them. These reasons lead to untreated mental disorders. As reported in the World Mental Health Survey, nearly 85% of serious mental health concerns in low- and middle-income countries (LMIC) did not receive any therapeutic management in the previous year [21]. Mental health screenings include questionnaires as well as one-on-one interviews. Many tools have been developed to evaluate anxiety and depression in children and adolescents. These assessment tools are often comprehensive, containing 20 to 80 questions each (Table 1.1).

Table 1.1: Common Mental Health Assessment Tools

Sr.no.	Assessment Tool	Psychological Disorder	Number of items	Reference
1.	DSM-5 Online Assessment Measures	Depression and Anxiety Disorders	25	[22]
2.	Patient Health Questionnaire (PHQ) Screeners	Anxiety, Depression, Eating Disorders	83	[23]
3.	Center for Epidemiological Studies Depression Scale for Children (CES-DC)	Depression Disorders	20	[24]

4.	Kutcher Adolescent Depression Scale (KADS)	Depression Disorders	6 to 11	[25], [26]
5.	Screen for Child Anxiety Related Disorders (SCARED)	Anxiety Disorders	41	[27]
6.	Spence Children's Anxiety Scale (SCAS)	Anxiety Disorders	35 to 45	[28]
7.	Spielberger State-Trait Anxiety Inventory for Children (STAI-C)	Non-disorder-specific Anxiety	20	[29]
8.	Revised Children's Manifest Anxiety Scale (RCMAS)	Anxiety Disorders	73	[30]
9.	Strengths and Difficulties Questionnaire (SDQ)	Anxiety, Depression, and Social Behaviors	25	[31]
10.	Fear Survey Schedule for Children-Revised (FSSC-R)	Fear	80	[32]

Another major issue is that with the present screening and assessment methods, there is a need for generalizability. The data used in epidemiological research on mental diseases predominantly comes from high-income nations, but to fully comprehend the actual global epidemiology, additional data from low- and middle-income countries (LMICs) are required. Most tools are revised and made with the needs of individuals in Europe or America in mind. The Revised Child Anxiety and Depression Scale (RCADS), for instance, was created by American researchers and was first tested on American individuals. Since then, it has been validated in different populations such as Australia [33], Denmark [34], Netherlands [35], Turkey [36], Ireland [37], El Salvador [38], and The United Kingdom [39]. As of yet, no study has been conducted in Pakistan. Sociodemographic factors of study populations are reported as key influencers in the development of early-onset psychological disorders [40]. Thus, to guarantee their effectiveness, methods of screening must be standardized and optimized with a focus on the Pakistani population as well.

The aforementioned information makes it evident that improvements are needed in the effectiveness and accessibility of mental health care services. This may be achieved by utilizing machine learning (ML) in developing intelligent clinical decision support systems that are driven by data. By assisting medical professionals in making evidence-based decisions, machine learning tools reduce the burden and improve patient care [41]. Furthermore, such user-friendly technologies' ease and adaptability can improve patient outcomes. In Chapter 2, the current state of ML integration in mental healthcare is covered in detail.

1.2 Depression and Anxiety

In the DSM-5, Depressive disorder is an umbrella term for illnesses that cause continuous feelings of sadness and accompanying changes that greatly impair one's capacity to function [42]. Among the depressive disorders is Major Depressive Disorder (MDD), which was previously placed in the “Mood Disorders” chapter of DSM-IV, and is now located in the “Depressive Disorder” section of DSM-5. Although this change may not seem like much, it has significant effects on the diagnosing process. A minor phrasing adjustment has broadened the diagnosis by adding hopelessness to the main mood requirement. For anxiety, DSM-5 removed obsessive-compulsive, acute stress, and post-traumatic stress disorders, providing a more precise and uniform definition of anxiety disorders. Separation anxiety disorder and selective mutism were classified as anxiety disorders rather than neurodevelopmental diseases, resulting in fewer differences between the childhood and adulthood categories of anxiety disorders [43]. Anxiety, in DSM-5, is defined as worrying about a potential threat. Anxiety disorders are defined as conditions characterized by overwhelming nervousness along with corresponding behavioral abnormalities. Fear and anxiety inevitably overlap, but they differ in the fact that fear is more often associated with an increase in neural activity necessary for fight-or-flight reactions or escape behaviors, in contrast, anxiety is more often associated with tense muscles, alertness in anticipation of danger, and cautious or avoidant behavior. Fear is an emotional response to a perceived or genuine impending threat, whereas anxiety is an anticipatory feeling of a threat in the future [44]. In primary healthcare settings, anxiety disorders are commonly diagnosed as panic disorder, social anxiety disorder, and

generalized anxiety disorder. Physical symptoms of anxiety disorders include rapid heartbeats, shortness of breath, and feeling lightheaded. Symptoms of anxiety disorders include nervousness, social fears, random or triggered panic attacks, worrying about the future, and avoidance behaviors [45].

1.3 Revised Child Anxiety and Depression Scale (RCADS)

The Revised Child Anxiety and Depression Scale (RCADS) is a revised version of the Spence Children's Anxiety Scale [46]. It is a freely available 47-item self-report measure used to evaluate children's symptoms that align with major depressive and anxiety disorders in the DSM-IV (see Appendix A). The RCADS has two versions: a long version, RCADS-47, with 47 items, and a short version, RCADS-25, with 25 items. The RCADS-47 is composed of 47 items and six subscales namely Major Depressive Disorder (MDD), Separation Anxiety Disorder (SAD), Social Phobia (SP), Generalized Anxiety Disorder (GAD), Panic Disorder (PD), Obsessive-Compulsive Disorder (OCD) (Table 1.3). Respondents must rate each of the 47 questions according to how frequently they can relate. A score of 0 for "Never," 1 for "Sometimes," 2 for "Often," and 3 for "Always" is assigned to each item.

There is also a Revised Child Anxiety and Depression Scale—Parent Version (RCADS-P), that is given to the parents or caregivers of the child to evaluate the child's anxiety and depression symptoms. The RCADS-P uses the same method for scoring as the RCADS finished by the child. Each subscale's responses are added up to provide a raw score, which is subsequently converted into a t-score. That t-score is then used to evaluate the child's depression and anxiety. The t-score conversion sheet is available on the RCADS website. The scoring may be done manually or automatically. For manual scoring, each item's score is assigned a number value of 0, 1, 2, or 3, representing Never, Sometimes, Often, and Always respectively. For each subscale, the numerical values of each item are summed together. The items that make up each subscale are mentioned in the Table 1.2. For example, to calculate Generalized Anxiety, the numerical values for items 1, 13, 22, 27, 35, and 37 are summed. Thus, the maximum possible score is 18, while the lowest is 0. Individuals between the ages of 8 and 18 can complete the RCADS, and parents or other

carers of young people can complete the RCADS-P. It takes five to ten minutes to complete the questionnaire. The copyright for the RCADS and all of its derivative works, including translations, belongs to Bruce F. Chorpita and Susan H. Spence. End users can download them from the RCADS website (<https://rcads.ucla.edu/>) [47].

Table 1.2: RCADS-47 Subscales and their Corresponding Items

Subscale	Corresponding Items	Score		
		MDD Score	Overall Depression Score	Total Internalizing Score
Major Depressive Disorder (MDD)	2, 6, 11, 15, 19, 21, 25, 29, 40, 47	MDD Score	Overall Depression Score	
Separation Anxiety Disorder (SAD)	5, 9, 17, 18, 33, 45, 46	SAD Score	Overall Anxiety Score	
Social Phobia (SP)	4, 7, 8, 12, 20, 30, 32, 38, 43	SP Score		
Generalized Anxiety Disorder (GAD)	1, 13, 22, 27, 35, 37	GAD Score		
Panic Disorder (PD)	3, 14, 24, 26, 28, 34, 36, 39, 41	PD Score		
Obsessive-Compulsive Disorder (OCD)	10,16, 23, 31, 42, 44	OCD Score		

1.4 Problem Statement

Childhood and adolescent mental health concerns are a neglected public health concern because of cultural stigma, a lack of awareness, and insufficient mental healthcare services. The Revised Child Anxiety and Depression Scale (RCADS) has established itself as a widely utilized self-report tool for diagnosing anxiety and depression symptoms in children and adolescents but this validation has been mostly carried out in Western

countries. Evaluating the RCADS in non-Western countries, particularly developing nations where anxiety and depression prevalence may differ, is also necessary.

Additionally, the current intervention options are expensive and time-consuming. Such constraints in mental healthcare increase the growing frequency of mental health problems. Failure to diagnose these issues early in childhood and teenage years can have major effects on the individual growing up. The improvement of existing screening methods for mental health concerns, particularly depression and anxiety using RCADS, may be achieved by employing machine learning techniques to create smart data-driven decision support systems. Scoring RCADS tests can be a tedious task for mental health professionals. It involves manually adding up all the answers and then looking up a table to convert the sum into a T-score. This process is time-consuming and error-prone. By training models directly on the raw RCADS responses, the need for manual calculations and table lookups can be eliminated. Moreover, machine learning can minimize human error and ensure consistent scoring. Incorporating machine learning into the RCADS scoring process has the potential to significantly improve efficiency, accuracy, and consistency and valuable time can be freed up for healthcare professionals. It can help healthcare providers by enabling them to quickly make well-informed judgments throughout the screening process.

1.5 Objectives

This study aims to achieve the following objectives:

- Analysis of each RCADS-47 item to identify significant features for the prediction of depression and anxiety using a Pakistani clinical sample.
- Generation of augmented data using the statistical distribution of the existing data to resolve the limitation of a small dataset.
- Developing a predictive model using machine learning methods to predict the disorder category based on RCADS scores.

1.6 Relevance to National Needs

In Pakistan, where resources for mental health are limited, managing psychological disorders is critical. With depression and anxiety affecting an estimated 15 million individuals of the population, as was mentioned in Section 1.1, limited services in remote areas make the issue worse [48]. Research and development in the mental health sector is much needed. Machine learning can be used for accurate screening instead of time-consuming and biased manual assessments. The data-driven decision-making can assist doctors in achieving accurate diagnoses. Furthermore, a machine learning model built on real-world data can promote early intervention by enabling accurate diagnosis and offering a deeper comprehension of the efficacy of RCADS in the Pakistani population. Healthcare developers can apply the suggested solution to create a convenient machine-learning tool.

The less time-consuming diagnosis and ultimately timely intervention for mental health concerns among children and adolescents will contribute to the Sustainable Development Goal (SDG) 3 - *Good Health and Well-being*. Additionally, the creation of a rapid screening tool will give mental health practitioners an efficient decision-support system, easing the strain of primary psychiatric consultations and accelerating the procedure. Enhancing and improving conventional healthcare methods can ultimately fulfill SDG 9 - *Industry, Innovation, and Infrastructure*

1.7 Thesis Structure

This dissertation adheres to a detailed framework to achieve the goals stated in Section 1.5. Chapter 2 outlines the literature assessment that was done to evaluate the extent of research on machine learning-driven prediction of mental health disorders in young people and to identify possible research gaps. The methodology of the study is elaborated upon in Chapter 3, and the results obtained are discussed in Chapter 4. Finally, a summary of the study, an acknowledgment of the limitations, and suggestions for addressing them are included in Chapter 5, which brings the dissertation to a close.

CHAPTER 02: LITERATURE REVIEW

Since self-report measures provide firsthand accounts of experiences not available from other sources, they are crucial for evaluating these disorders in children. Effective assessment methods are critical when dealing with children's mental health. It is essential to determine whether the techniques used to diagnose depression and anxiety disorders in kids are theoretically valid measures. RCADS has been demonstrated to be a valid tool for detecting and assessing anxiety and depression in both clinical and non-clinical populations of children and adolescents [49], [50]. RCADS has also demonstrated promising psychometric properties in different populations and cross-cultural research and meta-analyses have demonstrated the measure's strong psychometric qualities [51], [52]. Moreover, Over the last ten years, machine learning techniques have been included in healthcare systems for the diagnosis and likely prognosis of mental health disorders due to the need to discover efficient strategies for treating mental health disorders [53]. A machine learning algorithm is a computerized mechanism that, instead of being "hard coded"—that is, programmed to generate a certain result—uses input data to accomplish a specified job. These algorithms are "soft coded" meaning that they automatically change or adjust their design as a result of experience and repetition, hence improving their ability to do the intended objective. The algorithm then shapes itself most optimally so that it can generate the intended result from fresh, unknown data [54].

2.1 Evaluation of Psychometric Properties of RCADS

2.1.1 *Australia*

An Australian research team evaluated the psychometric properties of RCADS. There were 405 children and adolescents in the study, ranging in age from 6 to 18. The results displayed good test-retest reliability, internal consistency, and convergent validity using RCADS, along with robust structural validity and internal reliability. Preliminary data analysis covered various aspects, from assumption testing to confirmatory factor

analysis. The study identified RCADS as a valid instrument for diagnosing anxiety and depressive symptoms in Australian youth [33].

2.1.2 Turkey

In a research study with 483 participants aged 8 to 17, the psychometric properties of the Turkish version of RCADS-P were evaluated. The RCADS-P showed strong reliability and validity. The study confirmed the six-factor structure, aligned with DSM-based subscales, of the Turkish RCADS-P using confirmatory factor analysis. Notably, the instrument was able to distinguish between children with and without relevant diagnoses, which shows its discriminative validity. Gender differences occurred, with girls scoring higher on anxiety and depression. The results support the Turkish RCADS P as a useful tool for parents, emphasizing its usefulness in assessing current disorders [36].

2.1.3 Hawaii

In a study from 2005, the RCADS' psychometric qualities were analyzed in a clinical sample of 513 young people with an average age of 12.9 years and predominantly consisting of Caucasian, Hawaiian, Japanese American, Filipino, or multi-ethnic. Convergent and discriminant validity tests demonstrate that RCADS have favorable qualities when compared to clinical interviews and self-reports. Moreover, confirmatory factor analysis supported a valid factor structure, with all factor loadings being statistically significant. The study revealed gender and grade-level disparities in scores and emphasized the effectiveness of the RCADS in distinguishing between target and non-target groups for different diagnoses. The study found that the RCADS total anxiety scale did better than the usual methods in telling apart anxiety disorders from control cases [49].

In another study with 490 participants aged 6-18 from mental health clinics, RCADS-P displayed high internal consistency and strong discriminant validity between anxiety and depressive disorders. Both children and caregivers completed English questionnaires, and P ChIPS was used to conduct diagnostic interviews. RCADS showed good internal consistency at 0.95 and significant correlations were found with the CBCL Anxious/Depressed Syndrome Scale. The study confirms RCADS-P's discriminant

validity, utility in assessing internalizing problems, and specificity in measuring anxiety and depression [50].

Research in 2016 evaluated the shortened version of RCADS-P, the parent version. The research included a school samples that consisted of 967 children from grades 3 to 12 and a clinical sample that included 433 children ranging from 6 to 18 years old. For evaluating the psychometric properties, reliability, and validity of the RCADS-25-P, the study employed statistical analyses, including confirmatory factor analysis and reliability tests. According to the findings, the shorter version retained the reliable factor structure of the original 47-item version. Cronbach's alpha and test-retest correlations confirmed its consistency. The tool was able to differentiate among individuals with and without clinical diagnoses proved its discriminant validity. Examination of the parent-child agreement showed that the clinical sample had higher agreement than the school sample, possibly as a result of more severe symptoms. The 25-item RCADS version performed equally well, if not better, than the longer 47-item version, making it an appropriate alternative to diagnose anxiety and depression while reducing the diagnosis time [55].

2.1.4 Denmark

A Denmark research team aimed to evaluate the psychometric properties of the Danish version of RCADS. The sample consisted of 667 children aged 9 to 17 from community schools in Denmark. With a Cronbach's alpha of 0.96, the within-scale reliability was remarkable validating the RCADS DAN as a useful device for screening anxiety in young Danes. Anxiety and depression were observed at higher levels in girls than in boys, indicating a significant gender difference in total internalizing scores. A confirmatory factor analysis was also done, and the results confirmed the RCADS's 6-component structure. The study used RCADS and the Screen for Child Anxiety-Related Emotional Disorders (SCARED-R) for evaluation. The data analysis included reliability assessment and Pearson correlations for convergent validity. According to the findings, the RCADS DAN has similar psychometric features to those published in the United States and Europe, and it is a valid evaluation instrument for diagnosing anxiety and depression [34].

2.1.5 *Netherlands*

A study done in the Netherlands in 2015 aimed to investigate a school-based preventive program's effectiveness for childhood anxiety and depression in the Amsterdam area. The sample comprised 3636 children aged 8-13 from diverse ethnic backgrounds. It utilized RCADS and teacher reports to confirm the RCADS's reliability and validity for a multi-ethnic late childhood population. Confirmatory factor analysis replicated the RCADS's original factor structure. Good internal consistency and stability over three months were observed, supporting the RCADS's utility in identifying children in need of prevention programs. The study discovered a strong association between anxiety and Major Depressive Disorder (MDD) scales, implying that the MDD scale may better capture anxiety symptoms than depression symptoms. RCADS showed sensitivity to change and gender. Despite a small gap in detecting high-symptom children between RCADS and teacher reports, RCADS was recommended for screening and tracking changes in anxiety and depression symptoms over time. The study pointed out the need to use children's self-reports for successful screening in identifying and treating anxiety and depression symptoms [35].

Another study carried out in 2013 assessed the consistency of anxiety symptom measurement across adolescents using RCADS. The research was conducted as part of the Tracking Adolescents Individual Lives Survey (TRAILS), a Dutch cohort study, and included a sample of 2226 people from both urban and rural areas in the northern Netherlands. The ages of the participants ranged from 10 to 17 years old. The study focused on RCADS' longitudinal structure, examining its ability to consistently measure anxiety symptoms over time. The findings of the research concluded that RCADS effectively measures anxiety subtypes consistently across various time points in a general adolescent population. This consistency shows that changes in anxiety subscales are more likely to be true reflections of changes in anxiety levels rather than mistakes caused by measurement errors [56].

Another study conducted in 2020 aimed to assess the psychometric properties of two concise versions of the Revised Child Anxiety and Depression Scale (RCADS-25 and

RCADS-20) for screening anxiety and depression in school children and adolescents. The research involved 2,238 participants aged 8 to 18 from diverse primary and secondary schools in the Netherlands. The study evaluated the internal consistency, criterion validity, structural validity, test-retest reliability, and construct validity of the RCADS-25 and RCADS-20 scales. Key findings indicated that both scales demonstrated good structural validity, confirming a well-fitting four-factor model encompassing generalized anxiety disorder, separation anxiety disorder, social phobia, and major depressive disorder. Internal consistency was robust, with Cronbach's alpha coefficients ranging from 0.70 to 0.93. Test-retest reliability was favorable, exhibiting intra-class correlation coefficients between 0.70 and 0.91. Criterion validity was established through significant correlations with the Schedule for Affective Disorders and Schizophrenia for School-Age Children Present and Lifetime Version (K-SADS-PL). Additionally, construct validity was affirmed by significant correlations with other anxiety and depression measures (SCARED-NL and CDI-2). In summary, the study concludes that the RCADS-25 and RCADS-20 scales are both viable and efficient screening tools for anxiety and depression in school-age population [57].

2.1.6 United States of America

In one of the two studies, conducted by Chorpita et al, 1641 kids from various ethnic backgrounds were assessed for anxiety and depression using RCADS. Despite the anxiety items' satisfactory representation of the scale, a few of the items revealed weak correlations, suggesting issues with content validity. Study 2 involved 246 children aged around 12 years. A subset of 125 participants underwent a retest after a week, maintaining ethnic and age representativeness. RCADS reliability was assessed through a one-week test-retest, with alpha coefficients ranging from 0.71 to 0.85. The RCADS General Anxiety Disorder (GAD) scale demonstrated discriminant validity based on DSM-IV GAD criteria, emphasizing chronic worry over somatic symptoms. Interesting differences by sex and grade were also noted in RCADS subscales [46].

A research study aimed to preliminarily investigate the psychometric properties of the RCADS with a sample of children referred for possible ADHD, enhancing

generalizability to this population. The 117 participants included children from age 8 to 12 years old, predominantly male (65.8%), and white (76.9%). 83% of the participants met ADHD criteria. A six-factor RCADS structure was predicted by the research, along with strong discriminant, convergent, and reliability validity. Moderate degrees of anxiety and depression were demonstrated by the mean scale scores. Strong relationships were found when convergent validity was evaluated using one parent's report and the child's self-report measures. One child-self-report and two parent-reports were used to assess discriminant validity, which showed construct and criterion validity for both sets of reports. Internalizing symptoms were favorably correlated with RCADS scores, but externalizing behaviors were negatively correlated. Parent-child reports revealed moderate to strong connections with specific factor correlations. Adolescents who scored highly on the RCADS internalizing scale also showed noticeably higher anxiety levels. Moderate agreement was shown by the inter-correlations between the RCADS subscales, which ranged from 0.41 to 0.86, with a slightly lower correlation for depression [58].

Another research with a sample of 372 participants investigated the psychometric properties of the RCADS-P scale in children undergoing ADHD evaluations. The RCADS-P demonstrated consistent and accurate results, showing its validity and reliability. Analyses revealed that girls and children with internalizing disorders showed higher scores. The study investigated the prevalence of internalizing disorders among individuals with and without ADHD, revealing substantial differences in the prevalence rates for both groups. The RCADS-P showed good validity in assessing anxiety and depression when compared to Vanderbilt measures. The sensitivity-prioritizing ROC analysis demonstrated how well the RCADS-P screened for internalizing issues [59].

2.1.7 Ireland

In a different study that used statistical techniques like WLSMV and MIMIC, gender and age-based variations were the main focus of the assessment of the 47-item RCADS's psychometric qualities among Irish teenagers. 350 second-level students participated, the majority of whom were white (91.4%). A latent internalizing disorder factor was found to explain comorbidity, and the RCADS displayed strong internal

consistency, reliability, and convergent validity. Significantly greater levels of depression and anxiety were found in females. Distinct patterns in anxiety symptoms were shown by age-based disparities between school cycles, which suggests that the prevalence of anxiety disorders varies. The subscales measuring anxiety and depression showed moderate correlations, supporting the preliminary divergent validity [37].

2.1.8 United Kingdom

A study led by Karen McKenzie at the University of Cambridge in 2019, aimed to assess the validity of assessments designed for children and young people when applied to adults for measuring anxiety and depression. The Revised Children's Anxiety and Depression Scales (RCADS) were deemed reliable for adults, although adult responses primarily reflected general anxiety rather than specific anxiety disorder symptoms. Data for the original and short-form RCADS were obtained from 270 participants which consisted of 97 males (35.9%) ranging from 18 to 67 years old. An additional separate sample of 371 participants completed the 25-item short form RCADS items. The study explored the factorial structure of the RCADS in adults, favoring a bi-factor model to capture both general and specific anxiety subtypes. The findings highlighted that all RCADS versions provided reliable measures of general anxiety and depression in adults, encompassing various anxiety sub-dimensions. The study acknowledged potential sample bias, over-representing students and females, though existing research suggests that university students experience mental health issues at rates comparable to the general population [60].

2.1.9 Spain

Another research in 2021 examined the psychometric properties of the Spanish version of RCADS-25 with a sample of Salvadorian youth, and to establish its measurement invariance by gender. The sample for this study comprised 1296 3rd -12th grade students with a mean age of 12.73 years, of these, 716 identified as female (55.2%). The results of the study showed that the RCADS-25 had good internal consistency, test-retest reliability, and convergent validity with other measures of anxiety and depression.

The two-factor structure of the RCADS-25 was confirmed, and measurement invariance was established by gender. The study also provided normative data for the Salvadorian youth population across different age groups. The study concluded that the Spanish version of the RCADS-25 is a reliable and valid measure of anxiety and depression symptoms in Salvadorian youth, and can be used for school-based screening and clinical assessment purposes [38].

2.1.6 Cross-national

Another study investigated the cross-cultural measurement variabilities of RCADS in 3,908 African and White American children and adolescents aged 13-18. The study included 11 countries and tried to find the RCADS's validity among various cultural groups. The study used categorical confirmatory factor analysis (CCFA) and the multiple indicators multiple causes (MIMIC) model to analyze fit and differential item functioning (DIF) on a 47-item scale. The demographics showed significant variations. Cronbach's alpha demonstrated the RCADS's dependability, and CCFA revealed appropriate fit indexes for the majority of countries. Although eight items showed DIF, cross-cultural measurement invariance was established. Even after non-invariant elements were removed from the RCADS scores, statistically significant disparities between nations were found. The study concluded that the RCADS maintains its factor structure and validity across diverse cultural and ethnic groups [61].

Another study in 2022 delved into the cross-cultural evaluation of the RCADS in Spain, Chile, and Sweden, employing confirmatory factor analysis and multi-group CFA. Assessing three participant samples, the research spans school children and adolescents in non-clinical settings from the mentioned countries. The RCADS, known for robust psychometric properties, exhibited a superior fit to a unidimensional model in all countries, supporting its cross-cultural utility for gauging depression, anxiety, and obsessive-compulsive symptoms in youth. Internal consistency across RCADS subscales was consistently adequate to excellent. While the original six-factor model demonstrated satisfactory fit, modifications were suggested based on the Swedish sample. Latent means comparison, adjusted for age and gender differences, unveiled significant variations in

internalizing symptoms among Chilean, Spanish, and Swedish groups. Across all comparisons, girls consistently scored higher than boys on every factor. Strong evidence for measurement invariance across cultures was found, emphasizing the RCADS' consistency in diverse populations. The study underscores the importance of conducting measurement invariance tests in multinational research for cross-cultural reliability and advocates for further investigations into validity aspects across diverse cultural contexts [52].

2.2 Machine Learning for Depression and Anxiety

In machine learning, there are two major directions of pattern recognition; supervised and unsupervised learning. In supervised learning, the target classes of the data, that are being used to train the algorithm, are already labeled. Supervised pattern recognition uses labeled data to train a mapping function that connects input variable x to the output variable y . Unsupervised learning uses unlabeled data. Every case is assigned a label by the algorithm itself, by observing the data's underlying patterns. Unsupervised learning looks for structure and patterns in unlabeled data. Supervised machine learning includes regression and classification techniques like Random Forest, Decision Trees, K-Nearest Neighbors (KNN), and Logistic Regression. Unsupervised learning approaches have applications in clustering, dimensionality reduction, and anomaly detection [62]. Supervised machine learning techniques are more appropriate when it comes to ML-based mental health issue prediction as medical professionals are the only ones who should be making conclusions about an individual's health. Only to help with this process should machine learning techniques be employed.

A 2020 study employing machine learning algorithms in India predicted depression, anxiety, and stress. The DASS 21 questionnaire was used to gather data from 348 people, both employed and unemployed, from different socioeconomic backgrounds. Several machine learning algorithms were used for the classification process, including Catboost, Naive Bayes, Random Forest, Logistic Regression, and SVM. Of all the classifiers, Catboost produced the best levels of accuracy (82.6%) and precision (84.1%). In two data sets of 110 and 520 individuals, random forest produced the greatest accuracy

rate of 91% and 89% for stress, anxiety, and depression predictions. Because Random Forest was able to handle unbalanced classes, it was considered to be the best model, with Naive Bayes having the greatest accuracy. Despite its noteworthy contribution to psychological well-being, the study has many drawbacks. The accuracy of the classification algorithms is impacted by imbalanced classifications in the data. Because of the imbalance between the anxiety, depression, and stress classes, accuracy alone was insufficient to assess the models. Even though the study emphasized the significance of the f1 score in instances of classification imbalance, it also pointed out that all anxiety algorithms had low f1 scores, indicating difficulties in reliably predicting anxiety. Additionally, data from 348 employed and unemployed people aged 20 to 60 were gathered for the study. The sample size and individual participant characteristics may have limitations in representing demographics and variables that impact mental health [63].

Nemesure et al. predictively modeled anxiety and depression using a unique machine-learning technique. An ensemble of algorithmically different machine-learning techniques, including deep learning, were employed in the development of a unique machine-learning pipeline for this work. To enhance the performance of machine learning models, additional engineered features were created via feature engineering, such as Body Mass Index (BMI), Pulse Pressure, and Mean Arterial Pressure (MAP). The model's goal was to find significant indicators for GAD and MDD risk by training it to predict psychiatric disease using non-psychiatric input variables. 4,184 undergraduate students from the University of Nice Sophia Antipolis participated in the study; their ages and genders were evenly distributed. On the held-out test set, the ensemble model had a moderate predictive ability for the identification of GAD and MDD, with an AUC of 0.73 and 0.67 respectively. The study's reliance on a dataset of French college students poses a critical limitation, potentially impacting the generalizability of the findings despite the model's commendable predictive ability for psychiatric illnesses when non-psychiatric variables are used. The research sample showed a low prevalence of anxiety and depression, a typical disadvantage in mental health studies. In addition, the sample size was rather small [64].

Research in 2017 explored the potential of an automated system for predicting anxiety and depression in elderly patients to streamline diagnosis and referral using machine learning. The data was collected from 520 geriatric patients attending a government-operated tertiary care institution in Kolkata, India, and based on a thorough review of the literature and discussions with psychiatrists, eleven out of twenty features were chosen for predictive modeling using five different attribute evaluators (SU, CFS, PCA, GR, and OR). For binary classification, ten distinct classifiers—BN, logistic, RF, KS, RT, J48, SMO, MLP, RS, and NB—were used and evaluated for binary classification. The most suitable classifier was determined to be Random Forest (RF), which had a low false positive rate and 91% prediction accuracy. The chosen features and classifiers were externally validated using data from an additional 110 elderly patients. The research study discusses the difficulties in computing processes, specifically the slower processing of SVM. It is important to take into account the feasibility of using these models in healthcare environments that have limited resources [65].

In their study focused on the impact of depression and anxiety symptoms among schoolchildren, Qasrawi et al. conducted a comprehensive analysis using machine learning techniques to predict associated risk factors. 3,984 schoolchildren from public and refugee schools in the West Bank, ages 10 to 15, enrolled in fifth through ninth grades, were included in the study. Data was gathered using the Health Behaviors Schoolchildren questionnaire in the 2013–2014 academic year. Support vector machine (SVM), Random Forest (RF), neural network, decision tree, and Naive Bayes were the five machine learning approaches used in the study. SVM and Random Forest models showed the best accuracies in predicting both depression and anxiety symptoms with Random Forest obtaining 76.4% and 78.6% accuracy for depression and anxiety, and SVM reaching 92.5% and 92.4% accuracy respectively. The findings demonstrated the effectiveness of SVM and random forest in categorizing and predicting mental health issues in the student body under study. Key findings indicated that several factors such as school violence and bullying, domestic violence within the home environment, academic performance, and family income emerged as pivotal determinants that significantly influenced depression and anxiety levels among schoolchildren. Each of these factors was determined by the machine learning models to be a crucial determinant influencing student's mental well-being and cognitive

growth. The study emphasizes how early childhood anxiety and depression symptoms have a significant impact on mental health and cognitive development. It additionally emphasizes how machine learning techniques, in particular SVM and random forest models, hold promise for understanding and predicting the complex relations between risk factors. Although the suggested model's metrics are quite remarkable, the study's demographics raise uncertainty about the results' repeatability and generalizability. The children not only come from a certain ethnic group, but they also live in an area that is under occupation and unstable politically and socially. The subtleties of a child's well-being under such terrible conditions are not comparable to those of a child in an undisturbed, sovereign state [66].

2.3 Study Rationale

At the time of the literature review, there was no published research on the subject of predicting mental disorders among Pakistani children and adolescents; as a matter of fact, the vast majority of studies on RCADS had been conducted in the West. Since sociodemographic factors have a significant impact on the development of mental health issues, research on these areas is equally important in non-Western countries, especially developing nations where anxiety and depression prevalence may differ. This has been deemed an important research gap. Pakistan currently lacks a sufficient infrastructure for mental health treatment given the size of the country's population. To bridge the gap between the urgent need for mental health care and the availability of resources, effective smart screening procedures tailored for Pakistani adolescents are required. This study's use of local data from Pakistani children and teens will fill in these research gaps and have considerable effects on the multidisciplinary field of machine learning and mental health research.

CHAPTER 03: METHODOLOGY

A basic machine learning workflow, comprising data collection, pre-processing, feature selection, model training, and model assessment, was used in this study. The psychiatry department of Benazir Bhutto Hospital, Rawalpindi provided a secondary local dataset for this study. The data was pre-processed and scored using the SPSS batch code provided for RCADS. A collection of important features was chosen using statistical and computational techniques to train and test six machine learning algorithms that combined ensemble and classical methods. The performance of the models was evaluated using different metrics. Figure 3.1 provides a schematic diagram of the methodology.

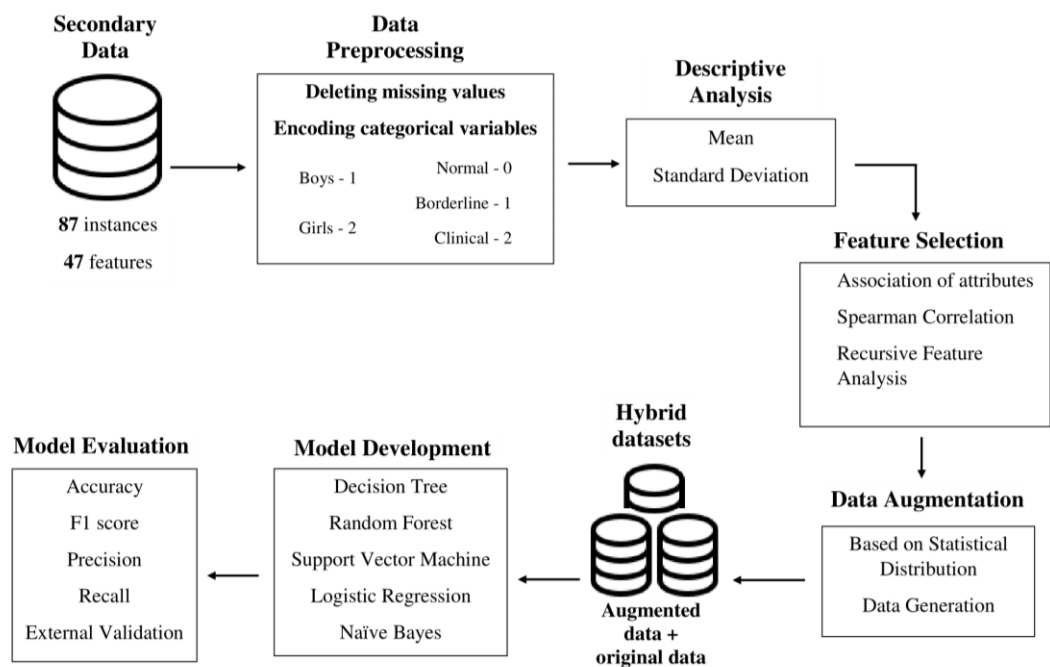


Figure 3.1: Workflow of the methodology

3.1 Data Collection

The data for the study was provided by the psychiatry department of Benazir Bhutto Hospital, Rawalpindi. Patients granted consent for the use of the data, understanding that it would be kept anonymous and used strictly for research. The data consisted of RCADS evaluations of 138 children and adolescents, 44 boys and 82 girls, ranging from grade 3 to grade 12. The scores of all individual items were calculated using the SPSS code available on the RCADS website to deduce the final evaluation of each individual. A t-score below the recommended cut-off point of 65 is in the normal range meaning that no referral to treatment is needed, unless clinical judgment suggests otherwise. A t-score between 65 and 69, is in the borderline clinical range, and whether or not referral is necessary needs to be clarified by doing a more comprehensive assessment or by using clinical judgment. A t-score of 70 or above is in the clinical range which indicates a referral to treatment is needed. The 47 items of the RCADS are divided into 6 subscales with each subscale having a set number of questions (Table 1.3). Cut-off scores for these subscales are the same.

3.2 Data Preprocessing

Data preprocessing is an important step in machine learning. If raw and unclean data is fed to a machine learning model, it affects the overall performance of the model. For 12 individuals, the gender was missing and for 46 individuals, the grade was missing. Since gender and grade are both vital pieces of information for evaluation, thus any cases that lacked either of this information were eliminated. Instances with even one missing value were deleted, leaving us with 87 instances, 34 boys and 53 girls. As stated in Section 3.1, a total score of less than 65 was categorized as Normal, a score between 65 and 69 as Borderline, and a score of 70 or more as Clinical. Of the 34 boys, 17 were categorized as normal, 3 as borderline, and 14 as clinical. Additionally, out of the 53 girls, 30 fell into the normal group, 4 into the borderline category, and 19 into the clinical category. For feature selection and machine learning, the 47 RCADS items were selected as input variables for predicting the overall evaluation.

3.3 Data Analysis

3.3.1 Descriptive Analysis

To describe the characteristics of the data, Kolmogorov-Smirnov and Chi-square tests of normality were used to evaluate the RCADS t-score of the total anxiety scale, the total internalizing scale, and each subscale to determine the distribution and skewness. The distribution was subsequently confirmed using skewness and kurtosis tests for each subscale. Testing for goodness-of-fit is a method used to assess how well a statistical model matches a collection of data. Single-sample Goodness-of-fit tests to check whether a sample could have been taken from a population with a specific distribution by taking into consideration a null and an alternative hypothesis. The Kolmogorov-Smirnov and Chi-square tests of normality are two examples of such tests that are used to determine whether or not the data fits a normal distribution. The test statistic is computed using the sample's empirical distribution function (EDF). If the p-value falls below the specified level of significance or the test statistic's value exceeds a critical value for a certain level of significance, the null hypothesis is rejected, indicating that the sample was taken from populations following a different distribution [67].

3.3.2 Internal Consistency and Reliability Analysis

Like previous studies, each subscale's internal consistency was measured using Cronbach's alpha to see how closely connected the RCADS items were. Cronbach's alpha, which is often represented by the lowercase Greek letter α , is a regularly used statistic to assess the internal consistency or reliability of rating scales. It quantifies the reliability of a score by calculating inter-item correlations among all items and the magnitude of Cronbach's alpha to summarize the information of questionnaire items [68]. It indicates how well the items of a questionnaire relate to one another. The alpha value ranges between 0 and 1. Higher numbers imply stronger internal consistency, indicating that the items are assessing the same underlying idea of the questionnaire [69]. Alpha values of 0.70 or higher are usually considered acceptable. For inter-item correlation, items with a correlation between 0.20 to 0.40 were regarded as sufficient [70]. The 47 items of the RCADS are

divided into 6 subscales with each subscale, however, in this study, only the overall internalizing evaluation was taken into account. None of the following steps made use of the evaluations from the other subscales.

3.4 Feature Selection

Feature selection is the step of reducing the number of input features before the development of a predictive model to improve the model's accuracy and efficiency. Two feature selection techniques, namely the filter method in which the selection of features is independent of a classifier and the wrapper method in which the features are selected using a classifier, were used to determine which of the 47 independent input variables were most relevant before machine learning algorithms were put into practice.

3.4.1 Chi-square Test of Independence

Often referred to as the Spearman Chi-square test, the chi-square test is one of the most used statistical methods for determining whether two categorical variables are associated or not [71]. It is a non-parametric (distribution-free) method for analyzing group differences when the dependent variable is nominal. The Chi-square is adaptable to the data's distribution, much like all other non-parametric statistics. In particular, homoscedasticity in the data or equality of variances among the groups is not required [72]. A single sample's two categorical variables may be tested for independence using the chi-square test to assess whether they are related to or independent of one another. Chi-square tests involve a calculation of a p-value and a test statistic (ρ). Rejection of the null hypothesis would suggest that there may be a relationship between one variable and another variable within the sample. Generally, values equal to 0.2 or below are deemed as weak associations [71]. In this study, C is calculated for each of the 47 RCADS items at a significance level of 0.05 using the Chi-square test.

3.4.2 Spearman Correlation

Correlation is the measurement of the relationship between two variables to determine whether they are unrelated, positively or negatively correlated, or neither.

Correlation coefficients are used to quantify the degree of relationship between variables. Stated differently, correlation coefficients quantify the degree of relationship, both direction and magnitude, between two variables. There are two types of correlation coefficients: positive and negative, i.e., the direction of the association, and high or low, i.e., the magnitude of the association. The Spearman's correlation coefficient, named after Charles Spearman, is a non-parametric measurement that uses ranks to measure the relationship between variables. It measures the degree to which a monotonic function can adequately explain the connection between two variables [73]. A correlation value of 0 denotes no association. Correlation coefficients range from -1 to +1, where -1 represents perfect negative correlation and +1 represents perfect positive correlation coefficients. In addition, correlation coefficients below 0.40 (positive or negative) are considered low, those between 0.40 and 0.60 are considered moderate, and those over 0.60 are considered strong. The statistical significance of the correlation is established by calculating a p-value in addition to r . P-values over the chosen threshold of significance denote significant findings, while those below it imply non-significant findings [74]. To identify significant features in this study, a correlation between the evaluation result and the 47 RCADS items was calculated. A correlation value of 0.4 and above was deemed satisfactory and values below 0.3 were considered poor [75].

3.4.3 Recursive Feature Elimination (RFE)

The Random Forest-Recursive Feature Elimination (RF-RFE) algorithm was used for the identification of significant features to be used during machine learning. The original idea of RFE was to train a model repeatedly, rank features, and then exclude the features with the lowest ranking to allow Support Vector Machines to do feature selection. Similar applications of this technique to Random Forests have shown that it works well when correlated features are present [77]. SVM's capacity to find strong predictors is impacted by the presence of correlated predictors, even if it supports non-linear connections between predictors. The Random Forest-Recursive Feature Elimination (RF-RFE) method is one proposed solution. Random Forest, a multiclass algorithm, has an intrinsic (unbiased) feature significance metric [76]. In this study, the 47 features (47 items

of RCADS) were tested iteratively using Python, and the features that yielded the highest performance metrics were chosen.

3.5 Data Augmentation

As stated in Section 3.2, 51 out of the 138 cases were discarded because of missing data, leaving us with 87 instances. 87 instances in a dataset are insufficient to create an efficient machine-learning model. Therefore, augmented data was generated utilizing the probability distribution followed by the data in the 47 items of RCADS. Programming language R and Rstudio were used for both confirming the probability distribution and generating the data.

3.5.1 Multinomial Probability Distribution

Multinomial distribution is a multidimensional generalization of the binomial distribution which is limited to only two possible outcomes (success and failure), to more than two values. The multinomial distribution is a distribution function for discrete processes, similar to the binomial distribution, but fixed probabilities are attached to each outcome. A multinomial distribution is a process that has k possible outcomes ($X_1, X_2, X_3, \dots, X_k$) with associated probabilities ($p_1, p_2, p_3, \dots, p_k$) such that $\sum p_i = 1$. The sum of the probabilities has to equal 1 because one of the results is certain to occur. Equation 3.5 below describes the density function of multinomial distribution [78].

$$(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1!x_2! \dots x_k!} p_1^{x_1} p_2^{x_2}, \dots, p_k^{x_k} \quad (3.1)$$

The original data's probability distribution was investigated to create augmented data that was as close to the original as possible. The multinomial distribution of the original data was investigated using the `chisq.test` function of the MASS package in R at an alpha level of 0.05. If the alpha value of each test is less than the critical value, it demonstrates that the applied distribution fits the data distribution.

3.5.2 Data Generation

To closely replicate the probability distribution of the original dataset, augmented data was generated utilizing the MASS and copula libraries in R. Five augmented datasets were created in varying sizes relative to the original data: 1:4 (four times the original size), 1:8 (eight times the original size), 1:12 (twelve times the original size), 1:16 (sixteen times the original size), and 1:20 (twenty times the original size).

3.6 Model Development

This study employs six machine learning models, including Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), and K-nearest Neighbor (KNN). These models were developed using Python and the Scikit-learn package, with the default parameters.

3.5.1 Decision Tree

A decision tree, a supervised machine learning method, is a graph to represent choices and their results in the form of a tree. It consists of a root node, internal nodes, and branches. The nodes represent an event or choice and the edges represent the decision rules or conditions. Each node represents attributes in a group that is to be classified and each branch represents a value that the node can take [79].

3.5.2 Random Forest Tree

The Random Forest is an ensemble method that makes use of several separate Decision Trees. There are three Random Forest hyper-parameters: the size of the nodes, the number of estimators, and the number of features that each node takes into factoring before splitting. Among commonly employed machine learning techniques, this regression tree method offers an additional level of model interpretability and prediction accuracy. With the use of random sampling and ensemble techniques, RF can provide better predictions and improved generalizations [80].

3.5.3 Support Vector Machine

In the last couple of years, the Support Vector Machine has gained immense popularity because of its ease of use and adaptability in handling various classification problems. An SVM's capacity to learn classification patterns with a balance between accuracy and repeatability is what gives it its effectiveness. While SVM is still rarely used for regression, it is now a popular method for classification with great adaptability that can be used in a variety of data science settings. The algorithm creates a reproducible hyperplane that optimizes the distance between the support vectors for the two class labels to train the model. A hyperplane is a line that separates data points into categories. The expected label for data that has not been shown yet may then be found using that hyperplane [81].

3.5.4 Logistic Regression

Logistic Regression predicts binary outcomes with two mutually exclusive states. Nonetheless, logistic regression has the capacity to account for several factors and allows the use of continuous or categorical predictors. The ratio of the chance of the event occurring divided by the probability of the event not occurring is the odds of the event of interest. The natural logarithm of the odds is used by the logistic regression model as a regression function of the predictors. An expansion of binary logistic regression that can classify multiple categories, like in this study, is multinomial logistic regression [82].

3.5.5 Naive Bayes

Naive Bayes is a Bayes Theorem-based classification algorithm that assumes predictors are independent. Simply put, a Naive Bayes classifier assumes that the features are unrelated. Even though this independence assumption is frequently violated in practice, the algorithm still provides excellent accuracy. Naive Bayes is primarily used for text classification. Naive Bayes utilizes sample data to estimate the posterior probability of each class y given x . These estimates are utilized for classification or other decision-support purposes [79].

3.5.6 K-Nearest Neighbour

The K-Nearest Neighbour is a relatively easy supervised machine learning approach that may be used for both regression and classification problems. Its main disadvantage is that, although it is simple to use and comprehend, it becomes noticeably slower as the volume of data being used increases [79].

3.7 Model Evaluation

A machine learning model's performance may be judged using a variety of metrics, including accuracy, precision, recall, and F1 score. These metrics track and evaluate the ML algorithms' performance quality during the training and testing stages and they do so by comparing the classification labels given by the model with the actual labels of the target in the dataset. The following metrics are used in this study to evaluate the model's performance.

3.6.1 Confusion Matrix

The confusion matrix albeit not a performance metric gives a tabular representation of the actual and predicted class labels. The rows in the table represent the actual class label in the data while the columns represent the class labels predicted by the ML model. The confusion matrix, in this study, is a 3x3 table since there are three class labels.

3.6.2 Accuracy

Accuracy is a performance metric that tells how many instances was the machine learning model able to predict correctly. It can be calculated by dividing the correct predictions by the total number of predictions.

$$Accuracy = \frac{\text{Number of correctly predicted instances}}{\text{Total predicted instances}} \quad (3.2)$$

3.6.3 Precision

Precision measures how many positive instances the machine learning model could predict correctly. It can be calculated by dividing the true positives i.e., actual positives in the data with the total number of positive predictions made by the model.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (3.3)$$

3.6.4 Recall

Recall is a performance metric that measures the number of correctly predicted positive classes from all the positive predicted classes in the data. It shows how many positive instances the model correctly identifies. It is also called sensitivity.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3.4)$$

3.6.6 F1 Score

The F1 score is the harmonic mean of precision and recall. It combines precision and recall into a single metric and is frequently used in binary and multi-class classification problems to better understand the model performance.

CHAPTER 04: RESULTS AND DISCUSSION

This chapter presents the findings of the suggested methodology and evaluates them in the context of related literature. The main objective of this research has been to create a clinical decision-support system utilizing RCADS for the early detection and treatment of two prevalent mental illnesses: anxiety and depression.

4.1 Descriptive Analysis

During the pre-processing, it was found that gender for 12 individuals and grade for 46 individuals were missing. Since gender and grade are both crucial pieces of information for the final evaluation, any case that lacked either information was discarded. Instances with even one missing value were deleted, leaving us with 87 instances, 34 boys (39%) and 53 girls (60%). As discussed in Section 3.1, a t-score of less than 65 was categorized as normal, a score between 65 and 69 as borderline, and a score of 70 or more as clinical. In the remaining dataset of 87 individuals, there were 47 normal cases (54%), 7 borderline cases (8%), and 33 clinical cases (37%). Among the 34 boys, 17 were

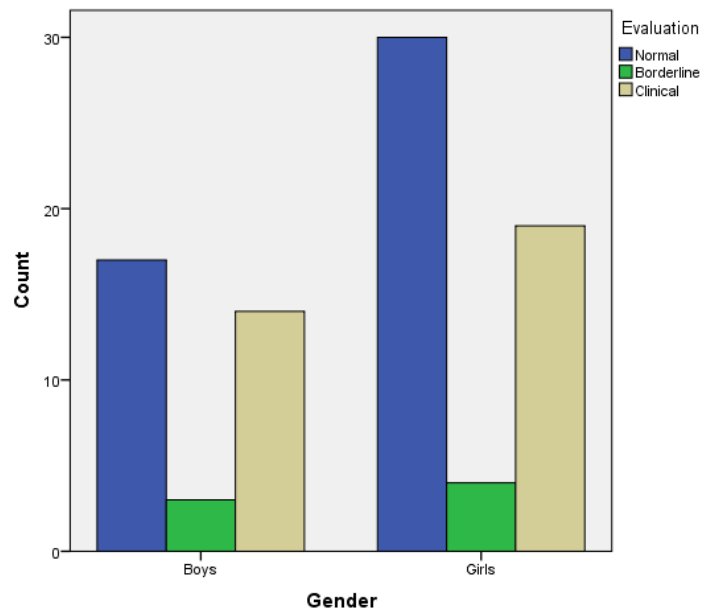


Figure 4.1: Bar chart of data distribution (gender)

classified as normal (19%), 3 as borderline (3%), and 14 as clinical (16%), and among the 53 girls, 30 (34.4%) were in the normal range, 4 (4.6%) in the borderline range, and 19 (21.8%) in the clinical range. Figure 4.1 and Figure 4.2 show bar charts of these distributions by gender and grade and the detailed distribution of the instances are shown in Table 4.1 and Table 4.2.

Table 4.1: Distribution of data (gender)

Gender	Evaluation			Total
	Normal	Borderline	Clinical	
	n (%)	n (%)	n (%)	
Boys	17 (19.5%)	3 (3.4%)	14 (16.1%)	34 (39.1%)
Girls	30 (34.4%)	4 (4.6%)	19 (21.8%)	53 (60.9%)
Total	47 (54%)	7 (8%)	33 (37.9%)	87

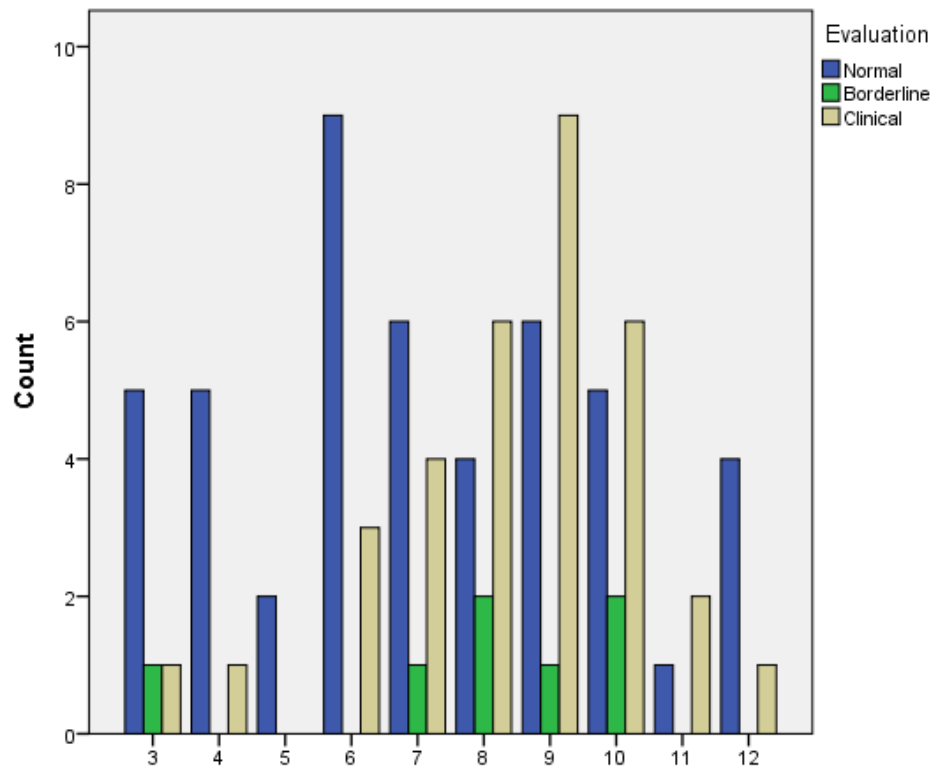


Figure 4.2: Bar chart of data distribution (grade)

Table 4.2: Distribution of data (grade)

Grade	Evaluation			Total
	Normal	Borderline	Clinical	
	n (%)	n (%)	n (%)	
3	5 (5.7%)	1 (1.1%)	1 (1.1%)	7 (8%)
4	5 (5.7%)	0 (0%)	1 (1.1%)	6 (6.8%)
5	2 (2.3%)	0 (0%)	0 (0%)	2 (2.3%)
6	9 (10.3%)	0 (0%)	3 (3.4%)	12 (13.7%)
7	6 (6.8%)	1 (1.1%)	4 (4.6%)	11 (12.6%)
8	4 (4.6%)	2 (2.3%)	6 (6.9%)	12 (13.7%)
9	6 (6.9%)	1 (1.1%)	9 (10.3%)	16 (18.3%)
10	5 (5.7%)	2 (2.3%)	6 (6.9%)	13 (14.9%)
11	1 (1.1%)	0 (0%)	2 (2.3%)	3 (3.4%)
12	4 (4.6%)	0 (0%)	1 (1.1%)	5 (5.7%)
Total	47 (54%)	7 (8%)	33 (37.9%)	87

Table 4.3 describes the normal, borderline, and clinical cases present in the data for the RCADS subscales; Major Depressive Disorder (MDD), Separation Anxiety Disorder (SAD), Social Phobia (SP), Generalized Anxiety Disorder (GAD), Panic Disorder (PD), Obsessive-Compulsive Disorder (OCD) as well as the Overall Anxiety Scale.

Table 4.3: Prevalence of the subscale disorders in the data

Subscale	Normal	Borderline	Clinical
	n (%)	n (%)	n (%)
Overall Anxiety Score	48 (55.2%)	10 (11.5%)	29 (33.3%)
Major Depressive Disorder (MDD)	52 (59.7%)	9 (10.3%)	26 (29.8%)
Generalized Anxiety Disorder (GAD)	71 (81.6%)	6 (6.9%)	10 (11.5%)
Obsessive-Compulsive Disorder (OCD)	56 (64.4%)	11 (12.6%)	20 (22.9%)
Panic Disorder (PD)	42 (48.3%)	9 (10.3%)	36 (41.4%)

Separation Anxiety Disorder (SAD)	34 (39.1%)	8 (9.2%)	45 (51.7%)
Social Phobia (SP)	67 (77%)	9 (10.3%)	11 (12.6%)

To describe the characteristics of the data, Kolmogorov-Smirnov and Chi-square tests of normality were used to evaluate the RCADS t-score of the total anxiety scale, the total internalizing scale, and each subscale to determine the distribution and skewness. The distribution was subsequently confirmed using skewness and kurtosis tests for each subscale. The t-scores of the overall internalizing scale (all six subscales), the overall anxiety scale (five subscales), and each subscale followed a normal distribution (p -value ≤ 0.05) as evident from the results of the Kolmogorov-Smirnov and Chi-Square tests of normality (Table 4.4). Skewness and kurtosis further verified the normal distribution (Table 4.5).

Table 4.4: Tests of normality of RCADS-47 subscales

Subscale	Mean	Std. dev	Kolmogorov-Smirnov		Chi-square	
			Statistics	Sig.	Statistics	Sig.
Overall Internalizing Score	63.43	18.87	0.08	0.58	5.33	0.50
Overall Anxiety Score	62.42	17.74	0.07	0.71	4.15	0.65
Major Depressive Disorder (MDD)	61.91	18.77	0.09	0.40	5.31	0.50
Generalized Anxiety Disorder (GAD)	51.75	13.16	0.11	0.15	6.93	0.32
Obsessive-Compulsive Disorder (OCD)	59.09	15.22	0.10	0.29	6.17	0.40
Panic Disorder (PD)	66.29	18.79	0.10	0.28	2.97	0.81
Separation Anxiety Disorder (SAD)	73.34	21.08	0.09	0.42	7.11	0.31
Std. dev: Standard Deviation						

Table 4.5: Skewness and Kurtosis of RCADS-47 subscales

Subscale	Range	Min	Max	Skewness	Kurtosis
Overall Internalizing Score	74	32	106	0.35	-0.61
Overall Anxiety Score	71	32	103	0.28	-0.63
Major Depressive Disorder (MDD)	73	32	105	0.46	-0.64
Generalized Anxiety Disorder (GAD)	51	27	78	0.35	-0.78
Obsessive-Compulsive Disorder (OCD)	68	32	100	0.46	-0.35
Panic Disorder (PD)	77	36	113	0.54	-0.45
Separation Anxiety Disorder (SAD)	89	37	126	0.18	-0.91
Social Phobia (SP)	56	28	84	0.32	-0.70

4.2 Internal Consistency and Reliability Analysis

The internal consistency of RCADS overall internalizing scale, overall anxiety scale, and each subscale was assessed using Cronbach's alpha. RCADS showed excellent internal consistency with an alpha of 0.953 (Table 4.6). The inter-item correlation was also computed for each item (Fig. 4.3). All 47 items combined had inter-item correlations that, on average, showed a weak to moderate relation, with most correlations falling between 0.1 and 0.6. This demonstrates that while the questionnaire's items are focused on one particular issue, they are well-diversified to avoid being redundant or repetitive. Removing items 3 and 5 resulted in a slight increase in the scale's internal consistency from 0.953 to 0.954. Conversely, removing items 10, 12, 15, 18, 19, 20, 21, 22, 23, 25, 29, 30, 31, 32, 34, 35, 37, 38, 39, 40, 41, 42, 44, and 45 caused a decrease from 0.953 to 0.952. Additionally, the removal of items 27 and 47 lowered the consistency from 0.953 to 0.951. However, these changes are too minor to be considered significant. Within each subscale, removing any item from the Major Depressive Disorder subscale, Obsessive-Compulsive

Disorder subscale, Separation Anxiety Disorder subscale, and Social Phobia subscale reduced their internal consistency. The removal of item 13 ("I worry that something awful will happen to someone in my family") increased the Generalized Anxiety Disorder subscale's internal consistency from 0.753 to 0.768. The subscale's internal consistency worsened when any other item was deleted. Similarly, when item 3 ("When I have a problem, I get a funny feeling in my stomach") was eliminated, the internal consistency of the Panic Disorder subscale improved slightly from 0.810 to 0.817.

Table 4.6: Internal consistency coefficient Cronbach's alpha for each subscale

Subscale	Cronbach's alpha
Overall Internalizing Scale	0.953
Overall Anxiety Scale	0.940
Major Depressive Disorder (MDD)	0.859
Generalized Anxiety Disorder (GAD)	0.753
Obsessive-Compulsive Disorder (OCD)	0.747
Panic Disorder (PD)	0.810
Separation Anxiety Disorder (SAD)	0.761
Social Phobia (SP)	0.835

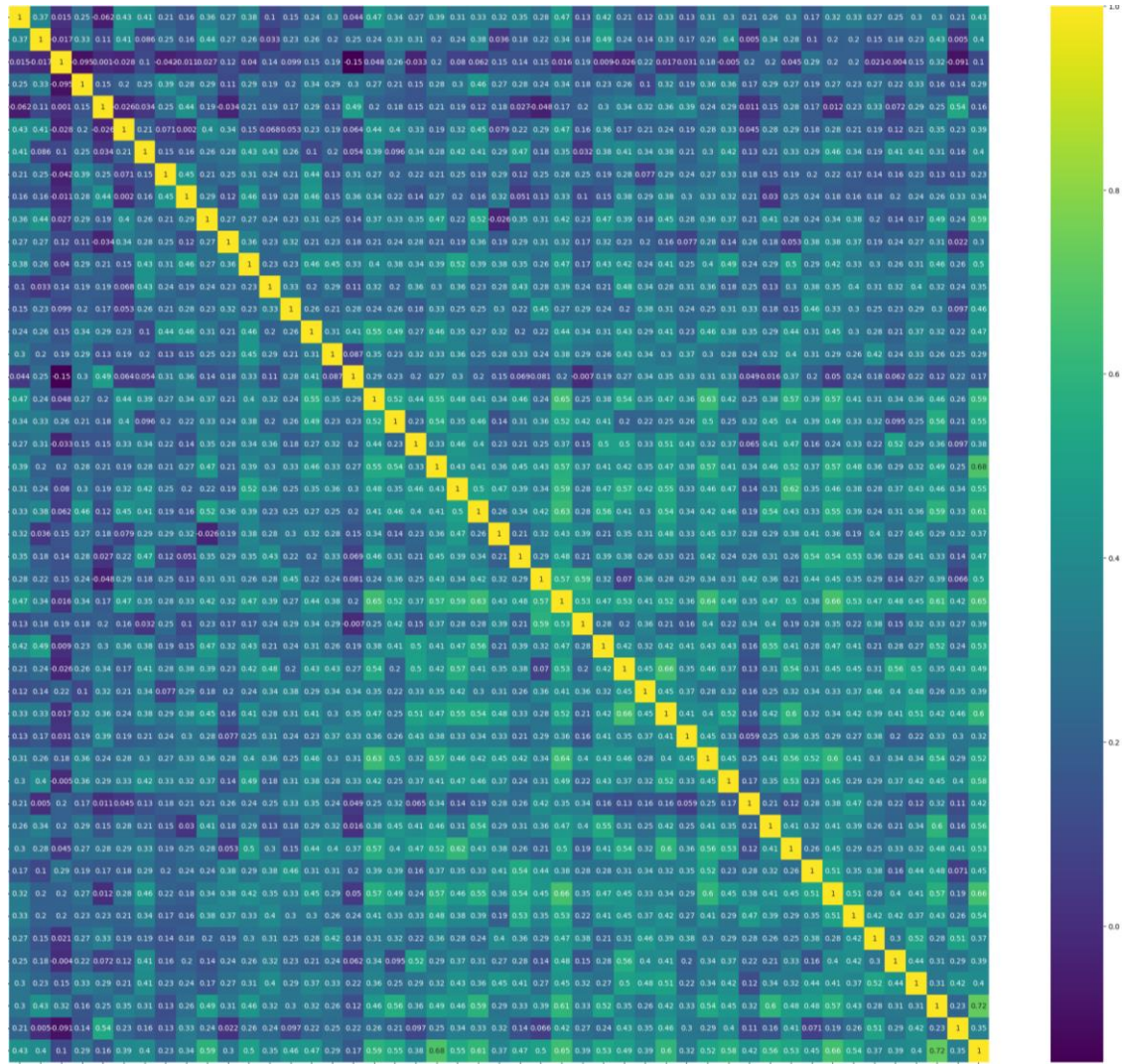


Figure 4.3: Heat map of inter-item correlation matrix

4.3 Feature Selection

Both filter and wrapper methods analysis revealed that majority of the features demonstrated a significant correlation with the target variable and played a crucial role for the final evaluation. The chi-square test of independence revealed that most of the features were statistically significant at an alpha level of 0.05 (Table 4.7).

Table 4.7: results of the Chi-square test of independence of the 47 features with the target

Feature	χ^2	Contingency Coefficient (C)	p-value
Rcads01	23.872	0.464	0.001
Rcads02	14.159	0.374	0.028
Rcads03	13.018	0.361	0.043
Rcads04	16.275	0.397	0.012
Rcads05	4.510	0.222	0.608
Rcads06	17.106	0.405	0.009
Rcads07	19.475	0.428	0.003
Rcads08	11.214	0.338	0.082
Rcads09	10.092	0.322	0.121
Rcads10	22.137	0.450	0.001
Rcads11	16.500	0.399	0.011
Rcads12	29.832	0.505	0.000
Rcads13	24.594	0.469	0.000
Rcads14	22.982	0.457	0.001
Rcads15	40.828	0.565	0.000
Rcads16	18.602	0.420	0.005
Rcads17	8.003	0.290	0.238
Rcads18	33.012	0.524	0.000
Rcads19	22.841	0.456	0.001
Rcads20	28.794	0.499	0.000
Rcads21	39.110	0.557	0.000
Rcads22	34.0.29	0.530	0.000
Rcads23	41.744	0.569	0.000
Rcads24	29.194	0.501	0.000
Rcads25	29.994	0.506	0.000
Rcads26	23.344	0.460	0.001
Rcads27	48.0.95	0.597	0.000

Rcads28	18.858	0.422	0.004
Rcads29	29.697	0.504	0.000
Rcads30	31.671	0.517	0.000
Rcads31	18.673	0.420	0.005
Rcads32	46.902	0.592	0.000
Rcads33	15.315	0.387	0.018
Rcads34	42.000	0.571	0.000
Rcads35	34.964	0.535	0.000
Rcads36	14.343	0.376	0.026
Rcads37	34.975	0.535	0.000
Rcads38	33.155	0.525	0.000
Rcads39	25.125	0.473	0.000
Rcads40	40.651	0.564	0.000
Rcads41	38.008	0.551	0.000
Rcads42	21.330	0.444	0.002
Rcads43	24.121	0.466	0.000
Rcads44	29.094	0.501	0.000
Rcads45	37.894	0.551	0.000
Rcads46	14.511	0.378	0.024
Rcads47	53.447	0.617	0.000
Feature with a weak association is highlighted in yellow.			
χ^2 : Chi-square test statistic, C: Contingency Coefficient of Chi-square test.			

Similarly, the correlation analysis between the target variable and the features showed that most features had a significant correlation. As stated in Section 3.4.2, correlations less than 0.3 were deemed weak. At the 0.05 alpha level, Rcads05 did not show significant correlation with the target variable. However, it is important to highlight that all correlations were significant at 0.01 alpha level (Table 4.8).

Table 4.8: results of Spearman correlation of the 47 features with the target

Feature	Spearman Correlation (ρ)	p-value
Rcads01	0.461	0.000
Rcads02	0.271	0.011
Rcads03	0.241	0.025
Rcads04	0.353	0.001
Rcads05	0.158	0.143
Rcads06	0.351	0.001
Rcads07	0.449	0.000
Rcads08	0.267	0.013
Rcads09	0.309	0.004
Rcads10	0.470	0.000
Rcads11	0.410	0.000
Rcads12	0.567	0.000
Rcads13	0.471	0.000
Rcads14	0.426	0.000
Rcads15	0.552	0.000
Rcads16	0.434	0.000
Rcads17	0.240	0.025
Rcads18	0.601	0.000
Rcads19	0.472	0.000
Rcads20	0.500	0.000
Rcads21	0.621	0.000
Rcads22	0.603	0.000
Rcads23	0.623	0.000
Rcads24	0.541	0.000
Rcads25	0.519	0.000
Rcads26	0.476	0.000
Rcads27	0.666	0.000
Rcads28	0.357	0.001

Rcads29	0.525	0.000
Rcads30	0.564	0.000
Rcads31	0.442	0.000
Rcads32	0.684	0.000
Rcads33	0.376	0.000
Rcads34	0.626	0.000
Rcads35	0.568	0.000
Rcads36	0.374	0.000
Rcads37	0.555	0.000
Rcads38	0.576	0.000
Rcads39	0.516	0.000
Rcads40	0.646	0.000
Rcads41	0.601	0.000
Rcads42	0.420	0.000
Rcads43	0.458	0.000
Rcads44	0.563	0.000
Rcads45	0.633	0.000
Rcads46	0.327	0.002
Rcads47	0.727	0.000
Feature with a weak correlation is highlighted in yellow. ρ: Spearman Correlation Coefficient		

RF-RFE selected 35 features as important to train a model. It eliminated Rcads02, Rcads03, Rcads05, Rcads06, Rcads09, Rcads10, Rcads17, Rcads19, Rcads28, Rcads33, Rcads36, and Rcads46 with an accuracy of 88%. RF-RFE starts with all features included in the model. After the initial run, the feature importance scores are calculated. The feature with the lowest importance score is then eliminated, and the model is run again. This process of calculating feature importance scores and eliminating the feature with the lowest score is repeated until either a stopping criterion is met or there is no further improvement in model accuracy. In this case, a specific number of features to be selected was not

provided; instead, the model determined it automatically (See Appendix B). The feature importance calculated during RF-RFE is represented as bar chart in Figures 4.4.

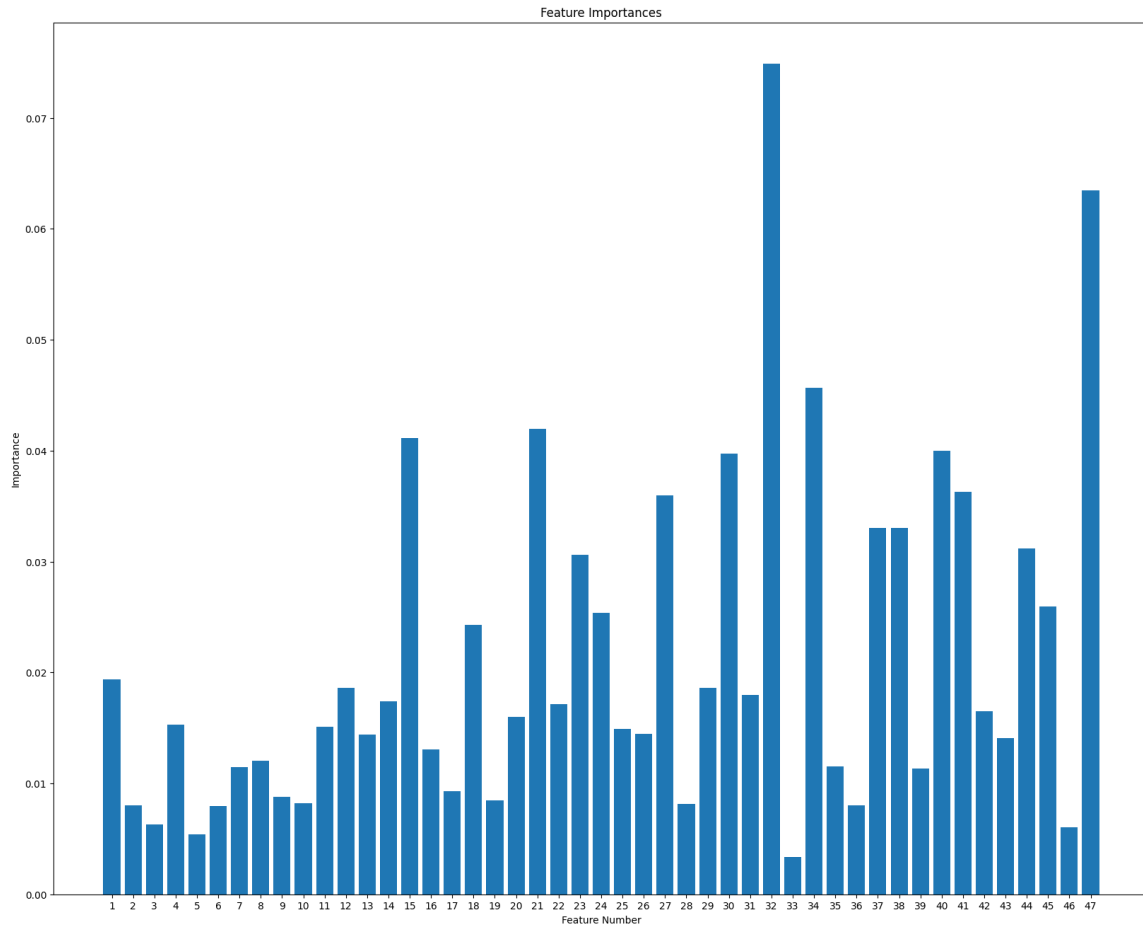


Figure 4.4: Bar chart of important features selected by RF-RFE

The number of features eliminated by the Chi-square test of independence, Spearman Correlation, and RF-RFE are 4, 1, and 12 respectively. One feature, Rcad05, was consistently identified as insignificant by all three methods. Because of this, it was removed from the data and not used during model training (Table 4.9). The elimination of this question is understandable as in Pakistani culture, joint families are common, and children are rarely left at home alone. While teenage boys may have some unsupervised time, it is less common for younger children and teenage girls. As a result, the concept of feeling afraid when alone at home is not a typical experience for most children in Pakistan.

This might explain why the particular question concerning this fear had no meaningful impact on predicting feelings of depression or anxiety in this context.

Table 4.9: Selected features from the three feature selection methods

Feature	Chi-Square Test of Independence	Spearman Correlation	RF-RFE
Rcads01	✓	✓	✓
Rcads02	✓	✓	✗
Rcads03	✓	✓	✗
Rcads04	✓	✓	✓
Rcads05	✗	✗	✗
Rcads06	✓	✓	✗
Rcads07	✓	✓	✓
Rcads08	✗	✓	✓
Rcads09	✗	✓	✗
Rcads10	✓	✓	✗
Rcads11	✓	✓	✓
Rcads12	✓	✓	✓
Rcads13	✓	✓	✓
Rcads14	✓	✓	✓
Rcads15	✓	✓	✓
Rcads16	✓	✓	✓
Rcads17	✗	✓	✗
Rcads18	✓	✓	✓
Rcads19	✓	✓	✗
Rcads20	✓	✓	✓
Rcads21	✓	✓	✓
Rcads22	✓	✓	✓

Rcads23	✓	✓	✓
Rcads24	✓	✓	✓
Rcads25	✓	✓	✓
Rcads26	✓	✓	✓
Rcads27	✓	✓	✓
Rcads28	✓	✓	✗
Rcads29	✓	✓	✓
Rcads30	✓	✓	✓
Rcads31	✓	✓	✓
Rcads32	✓	✓	✓
Rcads33	✓	✓	✗
Rcads34	✓	✓	✓
Rcads35	✓	✓	✓
Rcads36	✓	✓	✗
Rcads37	✓	✓	✓
Rcads38	✓	✓	✓
Rcads39	✓	✓	✓
Rcads40	✓	✓	✓
Rcads41	✓	✓	✓
Rcads42	✓	✓	✓
Rcads43	✓	✓	✓
Rcads44	✓	✓	✓
Rcads45	✓	✓	✓
Rcads46	✓	✓	✗
Rcads47	✓	✓	✓
Feature highlighted in yellow was common among all methods			

4.4 Data Augmentation

Due to the limited size of the original dataset for effective machine learning, data augmentation was chosen. The dataset consists of 47 items, each with 4 possible outcomes (0 = never, 1 = sometimes, 2 = often, and 3 = always). Upon reviewing the data, it appears to adhere to a multinomial distribution. To verify this, a chi-square test was conducted on each item, confirming the multinomial distribution (Table 4.10). The analysis utilized the MASS library in the R programming language to perform these tests. This probability distribution was used to generate 5 sets of data (1:4; four times the original data, 1:8; eight times the original data, 1:12; twelve times the original data, 1:16; sixteen times the original data, and 1:20; twenty times the original data) that mimic the distributional properties of the original 87 instances. Given that questions on the same subscale have a strong correlation with one another, data was generated for each subscale. This means that a child who responds "often" or "always" to one depression-related question is likely to respond similarly to other depression-related questions. To ensure the synthetic data accurately reflected these patterns and was not made up of random numbers, the average correlation between the questions and the target evaluation was calculated and used as input. Additionally, the probability of each possible answer (0, 1, 2, and 3) was also provided to the R code (See Appendix C and D). This approach was used to make sure the synthetic data produced results that were not just random values but closely mirrored the real data. The original 87 and the augmented instances were combined to generate a 'hybrid' dataset, which was then used in model development. Progressively expanding the dataset is based on well-established data augmentation and machine learning methods. In disciplines like deep learning, data augmentation is a common approach that involves artificially increasing the training dataset to improve the performance of the model [83]. Additionally, the concept of gradually expanding augmented data is consistent with ensemble methods like bootstrapping and bagging, which provide several data subsets to train various models [84].

Table 4.10: Results of the Chi-square test for multinomial distribution

RCADS	Chi-square	df	p-value	Null Hypothesis
Rcads01	0.0169	3	0.9994	Fail to reject
Rcads02	0.04023	3	0.9979	Fail to reject
Rcads03	0.036824	3	0.9981	Fail to reject
Rcads04	0.0275	3	0.9988	Fail to reject
Rcads05	0.016113	3	0.9995	Fail to reject
Rcads06	0.016021	3	0.9995	Fail to reject
Rcads07	0.012091	3	0.9996	Fail to reject
Rcads08	0.0054788	3	0.9999	Fail to reject
Rcads09	0.022784	3	0.9991	Fail to reject
Rcads10	0.03202	3	0.9985	Fail to reject
Rcads11	0.011431	3	0.9997	Fail to reject
Rcads12	0.0051168	3	0.9999	Fail to reject
Rcads13	0.0011789	3	0.9999	Fail to reject
Rcads14	0.0091817	3	0.9998	Fail to reject
Rcads15	0.023593	3	0.9990	Fail to reject
Rcads16	0.015973	3	0.9995	Fail to reject
Rcads17	0.017718	3	0.9994	Fail to reject
Rcads18	0.016254	3	0.9995	Fail to reject
Rcads19	0.0054946	3	0.9999	Fail to reject
Rcads20	0.0082746	3	0.9998	Fail to reject
Rcads21	0.01792	3	0.9994	Fail to reject
Rcads22	0.010762	3	0.9997	Fail to reject
Rcads23	0.0069459	3	0.9998	Fail to reject
Rcads24	0.0098667	3	0.9997	Fail to reject
Rcads25	0.0098667	3	0.9997	Fail to reject
Rcads26	0.0060208	3	0.9999	Fail to reject
Rcads27	0.031392	3	0.9985	Fail to reject
Rcads28	0.030651	3	0.9986	Fail to reject

Rcads29	0.0091817	3	0.9998	Fail to reject
Rcads30	0.010762	3	0.9997	Fail to reject
Rcads31	0.010914	3	0.9997	Fail to reject
Rcads32	0.0522	3	0.9969	Fail to reject
Rcads33	0.000346	3	0.9999	Fail to reject
Rcads34	0.018751	3	0.9993	Fail to reject
Rcads35	0.021728	3	0.9992	Fail to reject
Rcads36	0.011063	3	0.9997	Fail to reject
Rcads37	0.011137	3	0.9997	Fail to reject
Rcads38	0.15112	3	0.9851	Fail to reject
Rcads39	0.014231	3	0.9996	Fail to reject
Rcads40	0.0098101	3	0.9997	Fail to reject
Rcads41	0.010177	3	0.9997	Fail to reject
Rcads42	0.037352	3	0.9981	Fail to reject
Rcads43	0.11431	3	0.9997	Fail to reject
Rcads44	0.010983	3	0.9997	Fail to reject
Rcads45	0.000579	3	0.9999	Fail to reject
Rcads46	0.020549	3	0.9992	Fail to reject
Rcads47	0.017515	3	0.9994	Fail to reject

4.5 Model Development and Evaluation

First, the original dataset of 87 instances with all 47 features was used for model development. Code for the ML models is provided in Appendix E. For this dataset, RF correctly classified 45 out of 47 normal cases, 0 out of 7 borderline cases, and 29 out of 33 clinical cases. DT correctly classified 34 out of 47 normal cases, 2 out of 7 borderline cases, and 24 out of 33 clinical cases. SVM correctly classified 47 out of 47 normal cases, 0 out of 7 borderline cases, and 31 out of 33 clinical cases. LR correctly classified 46 out of 47 normal cases, 1 out of 7 borderline cases, and 32 out of 33 clinical cases. NB correctly classified 43 out of 47 normal cases, 3 out of 7 borderline cases, and 32 out of 33 clinical cases. KNN correctly classified 47 out of 47 normal cases, 2 out of 7 borderline cases, and

31 out of 33 clinical cases (Fig. 4.5A – F). The best-performing model was KNN with a 92% accuracy and 0.91 F1score (Table 4.11). While these accuracies are impressive, the machine learning models were likely over-fitted due to the insufficient training data of only 87 instances.

Table 4.11: Performance evaluation of the six ML models on the original dataset

Models	Accuracy	F1	Class	Precision	Recall
RF	0.85	0.81	Normal	0.87	0.96
			Borderline	0.00	0.00
			Clinical	0.83	0.88
DT	0.69	0.71	Normal	0.79	0.72
			Borderline	0.14	0.29
			Clinical	0.80	0.73
SVM	0.90	0.86	Normal	0.90	1.00
			Borderline	0.00	0.00
			Clinical	0.89	0.94
LR	0.91	0.89	Normal	0.94	0.98
			Borderline	0.33	0.14
			Clinical	0.91	0.97
NB	0.90	0.90	Normal	0.98	0.91
			Borderline	0.43	0.43
			Clinical	0.89	0.97
KNN	0.92	0.91	Normal	0.90	1.00
			Borderline	1.00	0.29
			Clinical	0.94	0.94

(A)	RF			(B)	DT			(C)	SVM		
	N	B	C		N	B	C		N	B	C
N	45	0	2	N	34	9	4	N	47	0	0
B	3	0	4	B	3	2	2	B	3	0	4
C	4	0	29	C	6	3	24	C	2	0	31

(D)	LR			(E)	NB			(F)	KNN		
	N	B	C		N	B	C		N	B	C
N	46	1	0	N	43	3	1	N	47	0	0
B	3	1	3	B	1	3	3	B	3	2	2
C	0	1	32	C	0	1	32	C	2	0	31

Figure 4.5: Confusion matrices of the six ML models on the original dataset

In the first of the five hybrid datasets, which was four times the size of the original dataset plus the original dataset (N=435). RF accurately classified 233/237 normal cases, 2/82 borderline cases, and 74/116 clinical cases. DT correctly classified 161/237 normal cases, 21/82 borderline cases, and 64/116 clinical cases. SVM correctly classified 224/237 normal cases, 13/82 borderline cases, and 87/116 clinical cases. LR correctly classified 211/237 normal cases, 28/82 borderline cases, and 87/116 clinical cases. NB correctly classified 197/237 normal cases, 42/82 borderline cases, and 83/116 clinical cases. And lastly, KNN correctly classified 227/237 normal cases, 7/82 borderline cases, and 73/116 clinical cases (Fig. 4.6A–F). Naive Bayes performed the best with 74% and 0.75 accuracy and F1 score respectively (Table 4.12).

Table 4.12: Performance evaluation of the six ML models on the 1:4 hybrid and feature-selected dataset

Models	Accuracy	F1	Class	Precision	Recall
RF	0.71	0.64	Normal	0.68	0.98
			Borderline	0.50	0.02
			Clinical	0.86	0.64
DT	0.57	0.57	Normal	0.75	0.68
			Borderline	0.23	0.26

			Clinical	0.49	0.55
SVM	0.74	0.71	Normal	0.78	0.95
			Borderline	0.36	0.16
			Clinical	0.78	0.75
LR	0.75	0.74	Normal	0.83	0.89
			Borderline	0.41	0.34
			Clinical	0.77	0.75
NB	0.74	0.75	Normal	0.88	0.83
			Borderline	0.38	0.51
			Clinical	0.81	0.72
KNN	0.71	0.66	Normal	0.72	0.96
			Borderline	0.29	0.09
			Clinical	0.78	0.63

(A)	RF		
	N	B	C
N	233	0	4
B	72	2	8
C	40	2	74

(B)	DT		
	N	B	C
N	161	42	35
B	30	21	31
C	24	28	64

(C)	SVM		
	N	B	C
N	224	7	6
B	51	13	18
C	13	16	87

(D)	LR		
	N	B	C
N	211	18	8
B	36	28	18
C	7	22	87

(E)	NB		
	N	B	C
N	197	36	4
B	25	42	15
C	1	32	83

(F)	KNN		
	N	B	C
N	227	4	6
B	60	7	15
C	30	13	73

Figure 4.6: Confusion matrices of the six ML models on the 1:4 hybrid dataset

In the second hybrid dataset, which was eight times the size of the original dataset plus the original dataset (N=783), RF accurately classified 435/444 normal cases, 3/136 borderline cases, and 124/203 clinical cases. DT correctly classified 303/444 normal cases, 37/136 borderline cases, and 98/203 clinical cases. SVM correctly classified 411/444

normal cases, 18/136 borderline cases, and 161/203 clinical cases. LR correctly classified 394/444 normal cases, 35/136 borderline cases, and 151/203 clinical cases. NB correctly classified 380/444 normal cases, 53/136 borderline cases, and 147/203 clinical cases. And lastly, KNN correctly classified 411/444 normal cases, 17/136 borderline cases, and 117/203 clinical cases (Fig. 4.7A – F). Naive Bayes performed the best with 74% accuracy and a 0.74 F1 score (Table 4.13).

Table 4.13: Performance evaluation of the six ML models on the 1:8 hybrid and feature-selected dataset

Models	Accuracy	F1	Class	Precision	Recall
RF	0.72	0.65	Normal	0.69	0.98
			Borderline	0.38	0.02
			Clinical	0.84	0.61
DT	0.56	0.57	Normal	0.71	0.68
			Borderline	0.23	0.27
			Clinical	0.51	0.48
SVM	0.75	0.71	Normal	0.78	0.93
			Borderline	0.45	0.13
			Clinical	0.74	0.79
LR	0.74	0.72	Normal	0.82	0.89
			Borderline	0.36	0.26
			Clinical	0.73	0.74
NB	0.74	0.74	Normal	0.87	0.86
			Borderline	0.35	0.39
			Clinical	0.75	0.72
KNN	0.70	0.66	Normal	0.71	0.93
			Borderline	0.30	0.12
			Clinical	0.78	0.58

(A)	RF			(B)	DT			(C)	SVM		
	N	B	C		N	B	C		N	B	C
N	435	2	7	N	303	78	63	N	411	10	23
B	116	3	17	B	66	37	33	B	83	18	35
C	76	3	124	C	57	48	98	C	30	12	161

(D)	LR			(E)	NB			(F)	KNN		
	N	B	C		N	B	C		N	B	C
N	394	29	21	N	380	48	16	N	411	18	15
B	66	35	35	B	50	53	33	B	101	17	18
C	18	34	151	C	6	50	147	C	65	21	117

Figure 4.7: Confusion matrices of the six ML models on the 1:8 hybrid dataset

In the third of the five hybrid datasets, which was twelve times the size of the original dataset plus the original dataset (N=1131). RF accurately classified 648/658 normal cases, 4/183 borderline cases, and 152/290 clinical cases. DT correctly classified 480/658 normal cases, 39/183 borderline cases, and 122/290 clinical cases. SVM correctly classified 612/658 normal cases, 25/183 borderline cases, and 228/290 clinical cases. LR correctly classified 594/658 normal cases, 52/183 borderline cases, and 227/290 clinical cases. NB correctly classified 587/658 normal cases, 51/183 borderline cases, and 222/290 clinical cases. And lastly, KNN correctly classified 616/658 normal cases, 15/183 borderline cases, and 140/290 clinical cases (Fig. 4.8A – F). Logistic Regression and Naive Bayes performed the best with Logistic Regression achieving 77% accuracy and a 0.75 F1 score and Naive Bayes achieving 76% accuracy and 0.75 F1 score (Table 4.14).

Table 4.14: Performance evaluation of the six ML models on the 1:12 hybrid and feature-selected dataset

Models	Accuracy	F1	Class	Precision	Recall
RF	0.71	0.64	Normal	0.69	0.98
			Borderline	0.57	0.02

			Clinical	0.80	0.52
DT	0.57	0.57	Normal	0.72	0.73
			Borderline	0.20	0.21
			Clinical	0.45	0.42
SVM	0.76	0.73	Normal	0.81	0.93
			Borderline	0.42	0.14
			Clinical	0.73	0.79
LR	0.77	0.75	Normal	0.85	0.90
			Borderline	0.40	0.28
			Clinical	0.75	0.78
NB	0.76	0.75	Normal	0.87	0.89
			Borderline	0.34	0.28
			Clinical	0.73	0.77
KNN	0.69	0.66	Normal	0.72	0.94
			Borderline	0.22	0.10
			Clinical	0.79	0.52

(A)	RF			(B)	DT			(C)	SVM		
	N	B	C		N	B	C		N	B	C
N	648	0	10	N	480	89	89	N	612	14	32
B	151	4	28	B	81	39	63	B	106	25	52
C	135	3	152	C	104	64	122	C	42	20	228

(D)	LR			(E)	NB			(F)	KNN		
	N	B	C		N	B	C		N	B	C
N	594	37	27	N	587	40	31	N	616	21	21
B	83	52	48	B	79	51	53	B	145	15	23
C	23	40	227	C	11	57	222	C	108	42	140

Figure 4.8: Confusion matrices of the six ML models on the 1:12 hybrid dataset

In the fourth hybrid dataset, which was sixteen times the size of the original dataset plus the original dataset (N=1479), RF accurately classified 819/845 normal cases, 4/270

borderline cases, and 207/364 clinical cases. DT correctly classified 586/845 normal cases, 66/270 borderline cases, and 169/364 clinical cases. SVM correctly classified 768/845 normal cases, 33/270 borderline cases, and 293/364 clinical cases. LR correctly classified 743/845 normal cases, 57/270 borderline cases, and 287/364 clinical cases. NB correctly classified 709/845 normal cases, 78/270 borderline cases, and 281/364 clinical cases. And lastly, KNN correctly classified 776/845 normal cases, 30/270 borderline cases, and 197/364 clinical cases (Fig. 4.9A – F). Again Logistic Regression and Naive Bayes performed the best with 73% and 72% accuracies and 0.72 F1 score respectively (Table 4.15).

Table 4.15: Performance evaluation of the six ML models on the 1:16 hybrid and feature-selected dataset

Models	Accuracy	F1	Class	Precision	Recall
RF	0.70	0.62	Normal	0.68	0.97
			Borderline	0.31	0.01
			Clinical	0.77	0.57
DT	0.69	0.71	Normal	0.72	0.69
			Borderline	0.22	0.24
			Clinical	0.44	0.44
SVM	0.74	0.70	Normal	0.78	0.91
			Borderline	0.35	0.12
			Clinical	0.73	0.80
LR	0.73	0.72	Normal	0.81	0.88
			Borderline	0.35	0.21
			Clinical	0.72	0.79
NB	0.72	0.72	Normal	0.84	0.84
			Borderline	0.33	0.29
			Clinical	0.71	0.77
KNN	0.68	0.64	Normal	0.71	0.92
			Borderline	0.26	0.11
			Clinical	0.73	0.54

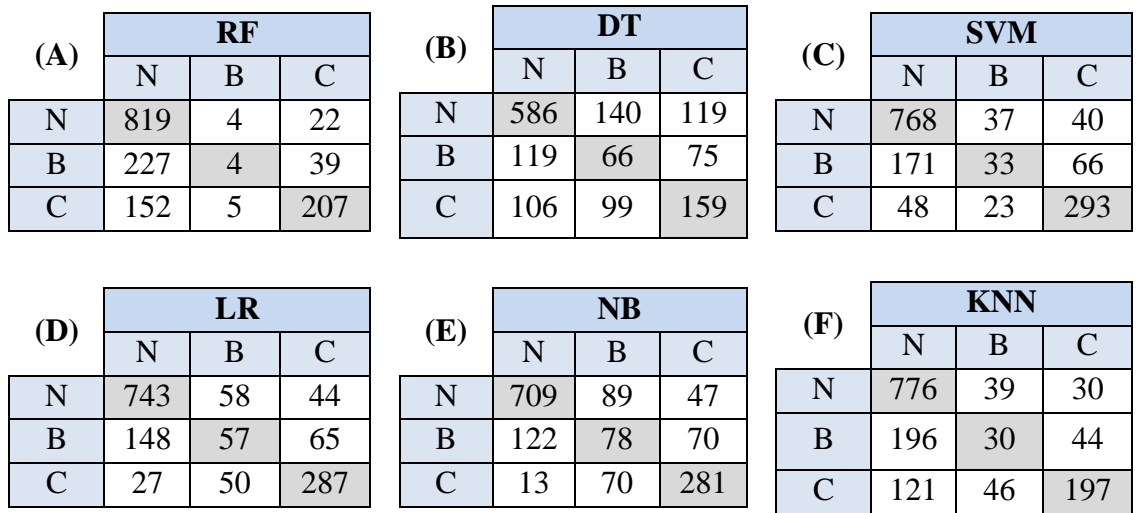


Figure 4.9: Confusion matrices of the six ML models on the 1:16 hybrid dataset

In the fifth and last hybrid dataset, which was twenty times the size of the original dataset plus the original dataset (N=1827), RF accurately classified 1035/1055 normal cases, 6/325 borderline cases, and 261/447 clinical cases. DT correctly classified 746/1055 normal cases, 85/325 borderline cases, and 195/447 clinical cases. SVM correctly classified 965/1055 normal cases, 63/325 borderline cases, and 340/447 clinical cases. LR correctly classified 956/1055 normal cases, 77/325 borderline cases, and 340/447 clinical cases. NB correctly classified 901/1055 normal cases, 126/325 borderline cases, and 341/447 clinical cases. And lastly, KNN correctly classified 987/1055 normal cases, 62/325 borderline cases, and 239/447 clinical cases (Fig. 4.10A – F). Naive Bayes performed the best achieving 75% accuracy and 0.75 F1 score (Table 4.16).

Table 4.16: Performance evaluation of the six ML models on the 1:20 hybrid and feature-selected dataset

Models	Accuracy	F1	Class	Precision	Recall
RF	0.71	0.64	Normal	0.69	0.98
			Borderline	0.32	0.02
			Clinical	0.84	0.58
DT	0.56	0.57	Normal	0.72	0.71

			Borderline	0.22	0.26
			Clinical	0.48	0.44
SVM	0.75	0.72	Normal	0.79	0.91
			Borderline	0.36	0.19
			Clinical	0.78	0.76
LR	0.76	0.74	Normal	0.83	0.91
			Borderline	0.35	0.25
			Clinical	0.77	0.79
NB	0.75	0.75	Normal	0.85	0.85
			Borderline	0.38	0.39
			Clinical	0.78	0.76
KNN	0.70	0.68	Normal	0.74	0.94
			Borderline	0.33	0.19
			Clinical	0.80	0.53

(A)	RF		
	N	B	C
N	1035	5	15
B	284	6	35
C	178	8	261

(B)	DT		
	N	B	C
N	746	176	133
B	158	85	82
C	133	119	195

(C)	SVM		
	N	B	C
N	965	54	36
B	200	63	62
C	49	58	340

(D)	LR		
	N	B	C
N	956	65	34
B	178	77	70
C	19	75	353

(E)	NB		
	N	B	C
N	901	118	36
B	140	126	59
C	14	92	341

(F)	KNN		
	N	B	C
N	987	43	25
B	228	62	35
C	126	82	239

Figure 4.10: Confusion matrices of the six ML models on the 1:20 hybrid dataset

The evaluation of machine learning models trained on the five sets of hybrid data revealed that most algorithms achieved accuracy exceeding 70%, except for the decision tree, which underperformed. Gradually increasing the data size led to a slight enhancement in the model's performance. However, the improvement was not substantial or drastic. While there was a noticeable positive trend, the overall impact on the model's accuracy and efficiency remained relatively modest. Among all models, Naive Bayes delivered the best overall performance (Table 4.17). The best-performing model was chosen based on the F1 score. In multiclass classification, choosing the optimal model based on the F1 score is helpful as it guarantees a balance between recall and precision, offering a thorough assessment of model performance [85]. Because it assigns equal weight to majority and minority classes, the F1 score provides a more informative measure across all classes, making it especially helpful for datasets with unequal class distributions. The F1 score makes model comparison easier by providing a single statistic that takes into account both false positives and false negatives. Because it may be weighted, macro, or averaged to represent performance across classes, it is well-suited for multiclass settings and ensures consistency [86]. Out of the three classes, the 'borderline' class was frequently observed to be falsely classified as the 'normal' class. The lack of borderline cases in the original data likely contributed to this class' poor representation since it is difficult for machine learning models to identify the underlying patterns in minority classes, which results in incorrect categorization.

Table 4.17: Summary of performance evaluations of the six ML models on five hybrid datasets

Models		1:4	1:8	1:12	1:16	1:20
RF	Accuracy	0.71	0.72	0.71	0.70	0.71
	F1 score	0.64	0.65	0.64	0.62	0.64
DT	Accuracy	0.57	0.56	0.57	0.55	0.56
	F1 score	0.57	0.57	0.57	0.55	0.57
SVM	Accuracy	0.74	0.75	0.76	0.74	0.85
	F1 score	0.71	0.71	0.73	0.70	0.72
LR	Accuracy	0.75	0.74	0.77	0.73	0.76

	F1 score	0.74	0.73	0.75	0.72	0.74
NB	Accuracy	0.74	0.74	0.76	0.72	0.75
	F1 score	0.75	0.74	0.75	0.72	0.75
KNN	Accuracy	0.71	0.70	0.69	0.68	0.70
	F1 score	0.66	0.66	0.66	0.64	0.68

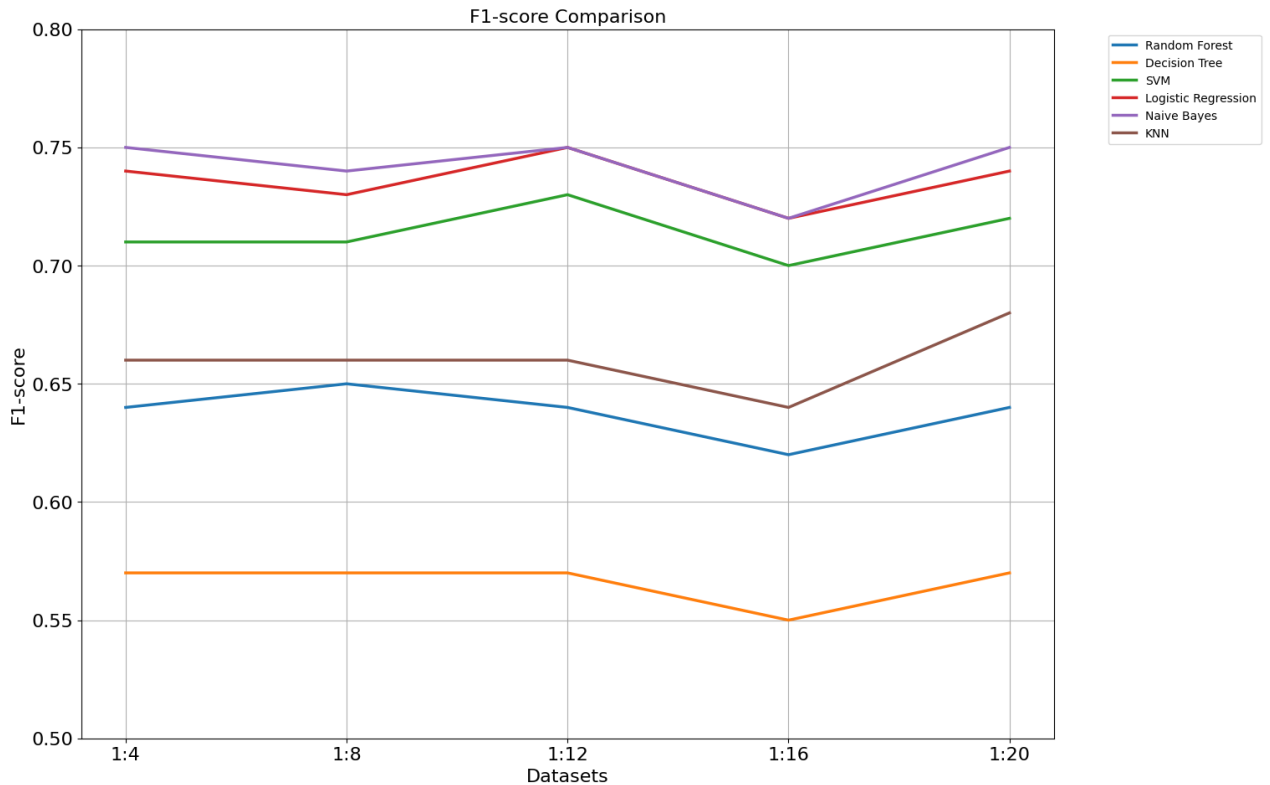


Figure 4.11: F1 score comparison of the six ML models across the five datasets

SVM, Logistic Regression, and Naive Bayes all achieved very similar accuracies, differing by only a few points. However, Naive Bayes demonstrated the best overall performance (Fig. 4.11) and was chosen as the best-performing model. External validation of Naive Bayes on a newly generated dataset of 500 samples resulted in an accuracy of 72% and an F1 score of 0.71, which are close to its training accuracy and F1 score. The Naive Bayes model outperformed other machine learning algorithms in the study for several reasons:

- **Independence Assumption:** Naive Bayes operates under the assumption that the features are independent given the class label. This assumption, while often unrealistic in real-world data, allows the model to simplify the computation of probabilities, making it efficient and effective, especially when the features are not highly correlated [87].
- **Categorical Data Suitability:** The model is particularly well-suited for categorical data, which aligns with the nature of the RCADS questionnaire responses (e.g., responses categorized as 0 = never, 1 = sometimes, 2 = often, 3 = always) [88].
- **Probability Estimates:** Naive Bayes provides probability estimates for each class, which helps in handling uncertainty and ambiguity in responses. This is particularly useful in psychological assessments where responses can be subjective [88].
- **Robustness to Irrelevant Features:** Naive Bayes can perform well even when some features are irrelevant, as it focuses on the conditional probabilities of the features given the class label.
- **Simplicity and Speed:** The simplicity of the Naive Bayes algorithm allows for quick training and prediction, making it a practical choice for applications requiring rapid assessments, such as mental health screenings.

These factors contributed to Naive Bayes achieving the best overall performance in the study, as indicated by its accuracy and F1 score during both the training and external validation phases.

CHAPTER 05: CONCLUSIONS AND FUTURE

RECOMMENDATIONS

5.1 Key Findings

The study's major findings show that ML-driven prediction models are effective in screening Pakistani children and adolescents for anxiety and depression with at least 70% accuracy. The problem of the limited sample size is addressed by hybrid data, which seems to be a viable substitute for real-world data. Feature selection reveals that Rcad5 (“I would feel afraid of being on my own at home”) does not have a strong correlation with depression and anxiety in the study population. Satisfactory performance has been observed by the SVM, Logistic Regression, and Naive Bayes algorithm, however, Naive Bayes had the best overall performance compared to the other algorithms. It implies that this algorithm seems to be an effective decision support system to help medical practitioners make well-informed screening decisions based on the chosen RCADS features. It is suggested to validate the study's findings with larger foreign data in the future and to develop the recommended method into a smart tool for end users. Nevertheless, the said tool cannot be utilized in clinical settings without more study and validation.

As far as we are aware, there has been no study like this in Pakistan. The current study is the first in terms of developing ML prediction models using Pakistani data and RCADS. In developing countries, where anxiety and depression prevalence may differ and where research is scarce and the burden of poor mental health is made worse by several issues like societal stigma, limited access to resources, and the high cost of mental health consultations, this preliminary contribution to the field of mental health can encourage more research and development concerning the integration of ML in healthcare practices.

5.2 Limitations

It is important to recognize the limitations of this study. The main limitation of the current study is the class imbalance in the sample. Although the small data size issue was

addressed by generating augmented data, it is important to acknowledge that this data, while helpful in increasing sample size, cannot eliminate the imbalances present in the original data. Since the biases in the original dataset persist in the augmented data as well. Therefore, an initial dataset with little to no class imbalance is recommended even when using data augmentation.

Moreover, only the total internalizing scale has been included in this study in the prediction models. The other RCADS subscales have not been included, as it would have been too complicated and outside the scope of the study to evaluate such an in-depth multiclass prediction model. However, these present findings can be further explored and developed in the future for the development of prediction models that include screening for the other subscales.

REFERENCES

- [1] K. J. Gilmore and P. Meersand, *Normal Child and Adolescent Development: A Psychodynamic Primer*. American Psychiatric Pub, 2013.
- [2] G. V. Polanczyk, G. A. Salum, L. S. Sugaya, A. Caye, and L. A. Rohde, “Annual Research Review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents,” *J. Child Psychol. Psychiatry*, vol. 56, no. 3, pp. 345–365, 2015, doi: 10.1111/jcpp.12381.
- [3] T. Potrebny, N. Wiium, and M. M.-I. Lundegård, “Temporal trends in adolescents’ self-reported psychosomatic health complaints from 1980-2016: A systematic review and meta-analysis,” *PLOS ONE*, vol. 12, no. 11, p. e0188374, Nov. 2017, doi: 10.1371/journal.pone.0188374.
- [4] F. C. Verhulst, J. van der Ende, R. F. Ferdinand, and M. C. Kasius, “The Prevalence of DSM-III-R Diagnoses in a National Sample of Dutch Adolescents,” *Arch. Gen. Psychiatry*, vol. 54, no. 4, pp. 329–336, Apr. 1997, doi: 10.1001/archpsyc.1997.01830160049008.
- [5] R. C. Kessler *et al.*, “Prevalence, Persistence, and Sociodemographic Correlates of DSM-IV Disorders in the National Comorbidity Survey Replication Adolescent Supplement,” *Arch. Gen. Psychiatry*, vol. 69, no. 4, pp. 372–380, Apr. 2012, doi: 10.1001/archgenpsychiatry.2011.160.
- [6] “Current Opinion in Psychiatry”, Accessed: Jun. 12, 2024. [Online]. Available: https://journals.lww.com/co-psychiatry/abstract/2007/07000/age_of_onset_of_mental_disorders__a_review_of.10.aspx
- [7] J. Kim-Cohen, A. Caspi, T. E. Moffitt, H. Harrington, B. J. Milne, and R. Poulton, “Prior Juvenile Diagnoses in Adults With Mental Disorder: Developmental Follow-Back

of a Prospective-Longitudinal Cohort,” *Arch. Gen. Psychiatry*, vol. 60, no. 7, pp. 709–717, Jul. 2003, doi: 10.1001/archpsyc.60.7.709.

[8] K. Beesdo-Baum and S. Knappe, “Developmental Epidemiology of Anxiety Disorders,” *Child Adolesc. Psychiatr. Clin.*, vol. 21, no. 3, pp. 457–478, Jul. 2012, doi: 10.1016/j.chc.2012.05.001.

[9] J. Ormel *et al.*, “Mental health in Dutch adolescents: a TRAILS report on prevalence, severity, age of onset, continuity and co-morbidity of DSM disorders,” *Psychol. Med.*, vol. 45, no. 2, pp. 345–360, Jan. 2015, doi: 10.1017/S0033291714001469.

[10] D. Johnson, G. Dupuis, J. Piche, Z. Clayborne, and I. Colman, “Adult mental health outcomes of adolescent depression: A systematic review,” *Depress. Anxiety*, vol. 35, no. 8, pp. 700–716, 2018, doi: 10.1002/da.22777.

[11] Z. M. Clayborne, M. Varin, and I. Colman, “Systematic Review and Meta-Analysis: Adolescent Depression and Long-Term Psychosocial Outcomes,” *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 58, no. 1, pp. 72–79, Jan. 2019, doi: 10.1016/j.jaac.2018.07.896.

[12] J. K. Kopinak, “Mental Health in Developing Countries: Challenges and Opportunities in Introducing Western Mental Health System in Uganda,” *Int. J. MCH AIDS*, vol. 3, no. 1, pp. 22–30, 2015.

[13] “Annual Report 2023 | UNICEF Pakistan.” Accessed: Jun. 11, 2024. [Online]. Available: <https://www.unicef.org/pakistan/reports/annual-report-2023>

[14] A. Javed, M. S. Khan, A. Nasar, and A. Rasheed, “Mental healthcare in Pakistan,” *Taiwan. J. Psychiatry*, vol. 34, no. 1, p. 6, 2020, doi: 10.4103/TPSY.TPSY_8_20.

[15] “Annual Report 2020 | UNICEF Pakistan.” Accessed: Jun. 11, 2024. [Online]. Available: <https://www.unicef.org/pakistan/reports/annual-report-2020>

- [16] A. Khalid, F. Qadir, S. W. Y. Chan, and M. Schwannauer, “Adolescents’ mental health and well-being in developing countries: a cross-sectional survey from Pakistan,” *J Ment Health*, vol. 28, no. 4, pp. 389–396, Jul. 2019, doi: 10.1080/09638237.2018.1521919.
- [17] T. A. Malik, S. Siddiqui, and A. Mahmood, “Behavioural and emotional problems among school children in Pakistan: A telephonic survey for prevalence and risk factors,” *J. Paediatr. Child Health*, vol. 55, no. 12, pp. 1414–1423, 2019, doi: 10.1111/jpc.14429.
- [18] S. Farooq, T. Yousaf, and S. Shahzad, “PREVALENCE OF EMOTIONAL AND BEHAVIOURAL PROBLEMS AMONG ADOLESCENTS IN PAKISTAN: A CROSS-SECTIONAL STUDY,” *J. Pak. Psychiatr. Soc.*, vol. 20, no. 01, Art. no. 01, Mar. 2023, doi: 10.63050/jpps.20.01.230.
- [19] “16BBF41.” Accessed: Jun. 11, 2024. [Online]. Available: <https://data.who.int/indicators/i/F08B4FD/16BBF41?m49=586>
- [20] R. Begum, F. R. Choudhry, T. M. Khan, F. S. Bakrin, Y. M. Al-Worafi, and K. Munawar, “Mental health literacy in Pakistan: a narrative review,” *Ment. Health Rev. J.*, vol. 25, no. 1, pp. 63–74, Jan. 2019, doi: 10.1108/MHRJ-08-2019-0026.
- [21] K. A. Ganasen, S. Parker, C. J. Hugo, D. J. Stein, R. A. Emsley, and S. Seedat, “Mental health literacy: focus on developing countries,” *Afr. J. Psychiatry*, vol. 11, no. 1, Art. no. 1, Apr. 2008, doi: 10.4314/ajpsy.v11i1.30251.
- [22] “Psychiatry.org - DSM-5-TR Online Assessment Measures.” Accessed: Jun. 11, 2024. [Online]. Available: <https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/assessment-measures>
- [23] “Patient Health Questionnaire (PHQ) Screeners. Free Download | phqscreeners.” Accessed: Jun. 11, 2024. [Online]. Available: <https://www.phqscreeners.com/select-screener>
- [24] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, “Center for Epidemiological Studies Depression Scale for Children (CES-DC),” in *STOP, THAT and*

One Hundred Other Sleep Scales, A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, Eds., New York, NY: Springer New York, 2011, pp. 93–96. doi: 10.1007/978-1-4419-9893-4_16.

[25] “6-ITEM Kutcher Adolescent Depression Scale: KADS-6,” 2008.

[26] “Kutcher Adolescent Depression Scale - 11-Item (KADS-11) - Psychology Tools.” Accessed: Jun. 11, 2024. [Online]. Available: <https://psychology-tools.com/test/kutcher-adolescent-depression-scale>

[27] “KSADS DSM 5 Supplement #3 | Child and Adolescent Bipolar Spectrum Services.” Accessed: Jun. 11, 2024. [Online]. Available: <https://pediatricbipolar.pitt.edu/resources/instruments>

[28] S. H. Spence, “Spence Children’s Anxiety Scale.” Jun. 11, 2012. doi: 10.1037/t10518-000.

[29] “Manual for the State-Trait Anxiety Inventory (Self-evaluation Questionnaire) | CiNii Research.” Accessed: Jun. 11, 2024. [Online]. Available: <https://cir.nii.ac.jp/crid/1370285712575158016>

[30] C. R. Reynolds and B. O. Richmond, “What I Think and Feel: A Revised Measure of Children’s Manifest Anxiety,” *J. Abnorm. Child Psychol.*, vol. 25, no. 1, pp. 15–20, Feb. 1997, doi: 10.1023/A:1025751206600.

[31] “Strengths & Difficulties Questionnaires (SDQ).” Accessed: Jun. 11, 2024. [Online]. Available: <https://www.sdqinfo.org/>

[32] T. H. Ollendick, “Reliability and validity of the revised fear survey schedule for children (FSSC-R),” *Behav. Res. Ther.*, vol. 21, no. 6, pp. 685–692, Jan. 1983, doi: 10.1016/0005-7967(83)90087-6.

[33] R. L. de Ross, E. Gullone, and B. F. Chorpita, “The Revised Child Anxiety and Depression Scale: A Psychometric Investigation with Australian Youth,” *Behav. Change*, vol. 19, no. 2, pp. 90–101, Jun. 2002, doi: 10.1375/beck.19.2.90.

- [34] B. H. Esbjørn, M. J. Sømhovd, C. Turnstedt, and M. L. Reinholdt-Dunne, “Assessing the Revised Child Anxiety and Depression Scale (RCADS) in a National Sample of Danish Youth Aged 8–16 Years,” *PLOS ONE*, vol. 7, no. 5, p. e37339, May 2012, doi: 10.1371/journal.pone.0037339.
- [35] M. P. Kösters, M. J. M. Chinapaw, M. Zwaanswijk, M. F. van der Wal, and H. M. Koot, “Structure, reliability, and validity of the revised child anxiety and depression scale (RCADS) in a multi-ethnic urban sample of Dutch children,” *BMC Psychiatry*, vol. 15, no. 1, p. 132, Jun. 2015, doi: 10.1186/s12888-015-0509-7.
- [36] V. Gormez *et al.*, “Psychometric Properties of the Parent Version of the Revised Child Anxiety and Depression Scale in a Clinical Sample of Turkish Children and Adolescents,” *Child Psychiatry Hum. Dev.*, vol. 48, no. 6, pp. 922–933, Dec. 2017, doi: 10.1007/s10578-017-0716-1.
- [37] A. Donnelly, A. Fitzgerald, M. Shevlin, and B. Dooley, “Investigating the psychometric properties of the revised child anxiety and depression scale (RCADS) in a non-clinical sample of Irish adolescents,” *J Ment Health*, vol. 28, no. 4, pp. 345–356, Jul. 2019, doi: 10.1080/09638237.2018.1437604.
- [38] J. Young, S. Ramachandran, R. Stewart, R. Orengo-Aguayo, and B. F. Chorpita, “Psychometric Properties of the Spanish Revised Child Anxiety and Depression Scale 25-Item Version in El Salvador,” *J. Psychopathol. Behav. Assess.*, vol. 43, no. 2, pp. 271–280, Jun. 2021, doi: 10.1007/s10862-020-09843-2.
- [39] I. Baron, R. Hurn, R. Adlington, E. Maguire, and L. Shapiro, “Revised Children’s Anxiety and Depression Scale (RCADS): Psychometric Properties in a Clinical Sample in the United Kingdom,” Jul. 2021, Accessed: Apr. 04, 2024. [Online]. Available: <http://uhra.herts.ac.uk/handle/2299/25861>
- [40] “Environmental Epidemiology”, Accessed: Jun. 11, 2024. [Online]. Available: https://journals.lww.com/environepidem/fulltext/2022/02000/early_environmental_quality_and_life_course_mental.2.aspx

- [41] “Thieme E-Journals - Yearbook of Medical Informatics / Full Text.” Accessed: Jun. 11, 2024. [Online]. Available: <https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0039-1677911>
- [42] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders: DSM-5*, 5th edition. Arlington, VA: American Psychiatric Association, 2013.
- [43] S.-C. Park and Y.-K. Kim, “Anxiety Disorders in the DSM-5: Changes, Controversies, and Future Directions,,” in *Anxiety Disorders*, vol. 1191, in *Advances in Experimental Medicine and Biology*, vol. 1191. , Springer, Singapore. Accessed: Jun. 12, 2024. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-32-9705-0_12#Abs1
- [44] R. Uher, J. L. Payne, B. Pavlova, and R. H. Perlis, “Major Depressive Disorder in Dsm-5: Implications for Clinical Practice and Research of Changes from Dsm-Iv,,” *Depress. Anxiety*, vol. 31, no. 6, pp. 459–471, 2014, doi: 10.1002/da.22217.
- [45] K. L. Szuhany and N. M. Simon, “Anxiety Disorders: A Review,,” *JAMA*, vol. 328, no. 24, pp. 2431–2445, Dec. 2022, doi: 10.1001/jama.2022.22744.
- [46] B. F. Chorpita, L. Yim, C. Moffitt, L. A. Umemoto, and S. E. Francis, “Assessment of symptoms of DSM-IV anxiety and depression in children: a revised child anxiety and depression scale,,” *Behav. Res. Ther.*, vol. 38, no. 8, pp. 835–855, Aug. 2000, doi: 10.1016/S0005-7967(99)00130-8.
- [47] “Home | Child FIRST – Focus on Innovation and Redesign in Systems and Treatment.” Accessed: Apr. 13, 2024. [Online]. Available: <https://rcads.ucla.edu/>
- [48] “WHO MiNDbank - WHO-AIMS Report on Mental Health System in Pakistan.” Accessed: Aug. 07, 2024. [Online]. Available: <https://extranet.who.int/mindbank/item/1305>
- [49] B. F. Chorpita, C. E. Moffitt, and J. Gray, “Psychometric properties of the Revised Child Anxiety and Depression Scale in a clinical sample,,” *Behav. Res. Ther.*, vol. 43, no. 3, pp. 309–322, Mar. 2005, doi: 10.1016/j.brat.2004.02.004.

- [50] C. Ebesutani, A. Bernstein, B. J. Nakamura, B. F. Chorpita, J. R. Weisz, and The Research Network on Youth Mental Health*, “A Psychometric Analysis of the Revised Child Anxiety and Depression Scale—Parent Version in a Clinical Sample,” *J. Abnorm. Child Psychol.*, vol. 38, no. 2, pp. 249–260, Feb. 2010, doi: 10.1007/s10802-009-9363-8.
- [51] J. A. Piqueras, M. Martín-Vivar, B. Sandin, C. San Luis, and D. Pineda, “The Revised Child Anxiety and Depression Scale: A systematic review and reliability generalization meta-analysis,” *J. Affect. Disord.*, vol. 218, pp. 153–169, Aug. 2017, doi: 10.1016/j.jad.2017.04.022.
- [52] M. Cervin, A. Veas, J. A. Piqueras, and A. E. Martínez-González, “A multi-group confirmatory factor analysis of the revised children’s anxiety and depression scale (RCADS) in Spain, Chile and Sweden,” *J. Affect. Disord.*, vol. 310, pp. 228–234, Aug. 2022, doi: 10.1016/j.jad.2022.05.031.
- [53] N. K. Iyortsuun, S.-H. Kim, M. Jhon, H.-J. Yang, and S. Pant, “A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis,” *Healthcare*, vol. 11, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/healthcare11030285.
- [54] I. El Naqa and M. J. Murphy, “What Is Machine Learning?” doi: 10.1007/978-3-319-18305-3_1.
- [55] “The Revised Child Anxiety and Depression Scale 25–Parent Version: Scale Development and Validation in a School-Based and Clinical Sample - Chad Ebesutani, Priya Korathu-Larson, Brad J. Nakamura, Charmaine Higa-McMillan, Bruce Chorpita, 2017.” Accessed: Apr. 04, 2024. [Online]. Available: <https://journals.sagepub.com/doi/abs/10.1177/1073191115627012>
- [56] C. M. Mathyssek, T. M. Olino, C. A. Hartman, J. Ormel, F. C. Verhulst, and F. V. A. Van Oort, “Does the Revised Child Anxiety and Depression Scale (RCADS) measure anxiety symptoms consistently across adolescence? The TRAILS study,” *Int. J. Methods Psychiatr. Res.*, vol. 22, no. 1, pp. 27–35, 2013, doi: 10.1002/mpr.1380.

- [57] L. Klaufus, E. Verlinden, M. van der Wal, M. Kösters, P. Cuijpers, and M. Chinapaw, “Psychometric evaluation of two short versions of the Revised Child Anxiety and Depression Scale,” *BMC Psychiatry*, vol. 20, no. 1, p. 47, Feb. 2020, doi: 10.1186/s12888-020-2444-5.
- [58] S. P. Becker, D. N. Schindler, A. S. Holdaway, L. Tamm, J. N. Epstein, and A. M. Luebbe, “The Revised Child Anxiety and Depression Scales (RCADS): Psychometric Evaluation in Children Evaluated for ADHD,” *J. Psychopathol. Behav. Assess.*, vol. 41, no. 1, pp. 93–106, Mar. 2019, doi: 10.1007/s10862-018-9702-6.
- [59] S. P. Becker, D. N. Schindler, A. M. Luebbe, L. Tamm, and J. N. Epstein, “Psychometric Validation of the Revised Child Anxiety and Depression Scales–Parent Version (RCADS-P) in Children Evaluated for ADHD,” *Assessment*, vol. 26, no. 5, pp. 811–824, Jul. 2019, doi: 10.1177/1073191117735886.
- [60] K. McKenzie, A. Murray, M. Freeston, K. Whelan, and J. Rodgers, “Validation of the Revised Children’s Anxiety and Depression Scales (RCADS) and RCADS short forms adapted for adults,” *J. Affect. Disord.*, vol. 245, pp. 200–204, Feb. 2019, doi: 10.1016/j.jad.2018.10.362.
- [61] D. Stevanovic *et al.*, “Cross-cultural measurement invariance of the Revised Child Anxiety and Depression Scale across 11 world-wide societies,” *Epidemiol. Psychiatr. Sci.*, vol. 26, no. 4, pp. 430–440, Aug. 2017, doi: 10.1017/S204579601600038X.
- [62] B. Rajoub, “Chapter 3 - Supervised and unsupervised learning,” in *Biomedical Signal Processing and Artificial Intelligence in Healthcare*, W. Zgallai, Ed., in *Developments in Biomedical Engineering and Bioelectronics.*, Academic Press, 2020, pp. 51–89. doi: 10.1016/B978-0-12-818946-7.00003-2.
- [63] A. Priya, S. Garg, and N. P. Tigga, “Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms,” *Procedia Comput. Sci.*, vol. 167, pp. 1258–1267, Jan. 2020, doi: 10.1016/j.procs.2020.03.442.

- [64] M. D. Nemesure, M. V. Heinz, R. Huang, and N. C. Jacobson, “Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence,” *Sci. Rep.*, vol. 11, no. 1, p. 1980, Jan. 2021, doi: 10.1038/s41598-021-81368-4.
- [65] A. Sau and I. Bhakta, “Predicting anxiety and depression in elderly patients using machine learning technology,” *Healthc. Technol. Lett.*, vol. 4, no. 6, pp. 238–243, 2017, doi: 10.1049/htl.2016.0096.
- [66] Q. R, V. P. Sp, A. A.-H. D, H. S, and A. Z, “Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis,” *PubMed*, 2022, Accessed: Jun. 15, 2024. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/35665695/>
- [67] B. M. Boyerinas, “Determining the Statistical Power of the Kolmogorov-Smirnov and Anderson-Darling Goodness-of-Fit Tests via Monte Carlo Simulation,” Defense Technical Information Center, Technical Report, Dec. 2016. Accessed: Jun. 21, 2024. [Online]. Available: <https://apps.dtic.mil/sti/citations/AD1029950>
- [68] A. Christmann and S. Van Aelst, “Robust estimation of Cronbach’s alpha,” *J. Multivar. Anal.*, vol. 97, no. 7, pp. 1660–1674, Aug. 2006, doi: 10.1016/j.jmva.2005.05.012.
- [69] M. Tavakol and R. Dennick, “Making sense of Cronbach’s alpha,” *Int. J. Med. Educ.*, vol. 2, pp. 53–55, Jun. 2011, doi: 10.5116/ijme.4dfb.8dfd.
- [70] R. L. Piedmont, “Inter-item Correlations,” in *Encyclopedia of Quality of Life and Well-Being Research*, A. C. Michalos, Ed., Dordrecht: Springer Netherlands, 2014, pp. 3303–3304. doi: 10.1007/978-94-007-0753-5_1493.
- [71] T. M. Franke, T. Ho, and C. A. Christie, “The Chi-Square Test,” <https://doi.org/10.1177/1098214011426594>, Nov. 2011, doi: 10.1177/1098214011426594.

- [72] M. L. McHugh, “The Chi-square test of independence”, Accessed: Jun. 21, 2024. [Online]. Available: <https://hrcak.srce.hr/clanak/152608>
- [73] A. Rebekić, Z. Lončarić, S. Petrović, and S. Marić, “PEARSON’S OR SPEARMAN’S CORRELATION COEFFICIENT - WHICH ONE TO USE?,” *Poljoprivreda*, vol. 21, no. 2, pp. 47–54, Dec. 2015, doi: 10.18047/poljo.21.2.8.
- [74] E. I. Obilor and E. Amadi, “Test for Significance of Pearson’s Correlation Coefficient (r),” Jan. 2018.
- [75] M. Mukaka, “A guide to appropriate use of Correlation coefficient in medical research,” *Malawi Med. J. J. Med. Assoc. Malawi*, vol. 24, no. 3, pp. 69–71, Sep. 2012.
- [76] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” *Chemom. Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, Sep. 2006, doi: 10.1016/j.chemolab.2006.01.007.
- [77] B. F. Darst, K. C. Malecki, and C. D. Engelman, “Using recursive feature elimination in random forest to account for correlated variables in high dimensional data,” *BMC Genet.*, vol. 19, no. 1, Art. no. 1, Sep. 2018, doi: 10.1186/s12863-018-0633-8.
- [78] “Multinomial Distribution - an overview | ScienceDirect Topics.” Accessed: Jun. 22, 2024. [Online]. Available: <https://www.sciencedirect.com/topics/mathematics/multinomial-distribution>
- [79] B. Mahesh, “Machine Learning Algorithms - A Review,” vol. 9, no. 1, 2018.
- [80] Y. Qi, “Random Forest for Bioinformatics,” 2012. doi: 10.1007/978-1-4419-9326-7_11.
- [81] D. A. Pisner and D. M. Schnyer, “Chapter 6 - Support vector machine,” in *Machine Learning*, A. Mechelli and S. Vieira, Eds., Academic Press, 2020, pp. 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7.

- [82] M. P. LaValley, “Logistic Regression,” *Circulation*, May 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.
- [83] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J. Big Data*, vol. 6, no. 1, Art. no. 1, Jul. 2019, doi: 10.1186/s40537-019-0197-0.
- [84] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, Art. no. 2, Aug. 1996, doi: 10.1007/BF00058655.
- [85] D. Powers, “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation,” *Mach Learn Technol*, vol. 2, Jan. 2008.
- [86] N. Japkowicz and S. Stephen, “The Class Imbalance Problem: A Systematic Study,” *Intell Data Anal*, vol. 6, pp. 429–449, Nov. 2002, doi: 10.3233/IDA-2002-6504.
- [87] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” *Mach. Learn.*, vol. 29, no. 2, pp. 131–163, Nov. 1997, doi: 10.1023/A:1007465528199.
- [88] F. T. Cruz, E. E. C. Flores, and S. J. C. Quispe, “Prediction of depression status in college students using a Naive Bayes classifier based machine learning model,” Jul. 25, 2023, *arXiv*: arXiv:2307.14371. doi: 10.48550/arXiv.2307.14371.

**APPENDIX A: THE REVISED CHILD ANXIETY AND
DEPRESSION SCALE (RCADS-47)**

Item	Question
Rcads01	I worry about things
Rcads02	I feel sad or empty
Rcads03	When I have a problem, I get a funny feeling in my stomach
Rcads04	I worry when I think I have done poorly at something
Rcads05	I would feel afraid of being on my own at home
Rcads06	Nothing is much fun anymore
Rcads07	I feel scared when I have to take a test
Rcads08	I feel worried when I think someone is angry with me
Rcads09	I worry about being away from my parents
Rcads10	I get bothered by bad or silly thoughts or pictures in my mind
Rcads11	I have trouble sleeping
Rcads12	I worry that I will do badly at my school work
Rcads13	I worry that something awful will happen to someone in my family
Rcads14	I suddenly feel as if I can't breathe when there is no reason for this
Rcads15	I have problems with my appetite
Rcads16	I have to keep checking that I have done things right (like the switch is off, or the door is locked)
Rcads17	I feel scared if I have to sleep on my own
Rcads18	I have trouble going to school in the mornings because I feel nervous or afraid
Rcads19	I have no energy for things
Rcads20	I worry I might look foolish
Rcads21	I am tired a lot
Rcads22	I worry that bad things will happen to me
Rcads23	I can't seem to get bad or silly thoughts out of my head

Rcads24	When I have a problem, my heart beats really fast
Rcads25	I cannot think clearly
Rcads26	I suddenly start to tremble or shake when there is no reason for this
Rcads27	I worry that something bad will happen to me
Rcads28	When I have a problem, I feel shaky
Rcads29	I feel worthless
Rcads30	I worry about making mistakes
Rcads31	I have to think of special thoughts (like numbers or words) to stop bad things from happening
Rcads32	I worry what other people think of me
Rcads33	I am afraid of being in crowded places (like shopping centers, the movies, buses, busy playgrounds)
Rcads34	All of a sudden I feel really scared for no reason at all
Rcads35	I worry about what is going to happen
Rcads36	I suddenly become dizzy or faint when there is no reason for this
Rcads37	I think about death
Rcads38	I feel afraid if I have to talk in front of my class
Rcads39	My heart suddenly starts to beat too quickly for no reason
Rcads40	I feel like I don't want to move
Rcads41	I worry that I will suddenly get a scared feeling when there is nothing to be afraid of
Rcads42	I have to do some things over and over again (like washing my hands, cleaning, or putting things in a certain order)
Rcads43	I feel afraid that I will make a fool of myself in front of people
Rcads44	I have to do some things in just the right way to stop bad things from happening
Rcads45	I worry when I go to bed at night
Rcads46	I would feel scared if I had to stay away from home overnight
Rcads47	I feel restless

APPENDIX B: PYTHON CODE FOR RF-RFE

```
# import packages

from google.colab import files

import pandas as pd

# Uploading the CSV file from computer

uploaded = files.upload()

filename = 'data.csv'

df = pd.read_csv(filename)

print(df)

# Separate the features and target variables

X = df.drop('Target', axis=1)

y = df['Target']

import numpy as np

import matplotlib.pyplot as plt

from sklearn.datasets import make_classification

from sklearn.model_selection import cross_val_score,
StratifiedKFold

from sklearn.feature_selection import RFE

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier
```



```

from sklearn.pipeline import Pipeline

# Function to evaluate a model using stratified 5-fold cross-
validation

def evaluate_model(model, X, y):

    cv = StratifiedKFold(n_splits=5, shuffle=True,
random_state=42)

    scores = cross_val_score(model, X, y, scoring='accuracy',
cv=cv, n_jobs=-1)

    return np.mean(scores)

# Initialize a RandomForestClassifier as the estimator

estimator = RandomForestClassifier(n_estimators=100,
random_state=42)

# Initialize a list to store the mean accuracies for each number
of selected features

mean_accuracies = []

# Initialize a list to store the names of the optimal number of
selected features

optimal_feature_counts = []

# Loop through different numbers of selected features

for i in range(1, 48):

    rfe = RFE(estimator=estimator, n_features_to_select=i)

    model = RandomForestClassifier()

    pipeline = Pipeline(steps=[('s', rfe), ('m', model)])

```

```

    mean_accuracy = evaluate_model(pipeline, X, y)
    mean_accuracies.append(mean_accuracy)

    optimal_feature_counts.append(i)

    print(f"Number of Selected Features: {i}, Mean Accuracy:
    {mean_accuracy:.3f}")

# Find the index of the maximum mean accuracy
optimal_idx = np.argmax(mean_accuracies)
optimal_features = optimal_feature_counts[optimal_idx]
print(f"Optimal Number of Selected Features: {optimal_features}")

# Plot feature importances using the RandomForestClassifier
estimator.fit(X, y)
importances = estimator.feature_importances_
print(importances)

plt.figure(figsize=(20, 16))

plt.title("Feature Importances")

plt.bar(range(len(importances)), importances,
        tick_label=np.arange(1, len(importances) + 1))

plt.xlabel("Feature Number")

plt.ylabel("Importance")

plt.show()

```

APPENDIX C: R CODE FOR CHI-SQUARE TEST FOR MULTINOMIAL DISTRIBUTION

```
# Load libraries

library(MASS) # for chisq.test

# Read data

data=read.csv("RCADS.csv",header=TRUE)

# Check for missing values (not recommended for multinomial)
if (any(is.na(data$rcads))) {
  stop("Data contains missing values. Multinomial not suitable!")
}

# Get observed counts

observed <- table(data$rcads)

# Define total number of trials (observations)

n <- sum(observed)

# Define the vector of unequal probabilities (replace with your
values)

p <- c(0.08, 0.07, 0.02, 0.14, 0.13, 0.14, 0.18, 0.15, 0.03,
0.06) # Probabilities for categories 0, 1, 2, 3

# Check if probability vector length matches category count

if (length(p) != nrow(table(data$rcads))) {
  stop("Probability vector length must match number of
categories!")
}
```

```

# Calculate expected counts using the probabilities
expected <- n * p

# Create table for observed and expected counts

counts_table <- cbind(Category = names(observed), Observed =
observed, Expected = expected)

# Print the table

cat("Table of Observed and Expected Counts:\n")

print(counts_table)

# Perform Chi-squared goodness-of-fit test

chisq.result <- chisq.test(observed, p = expected/n)

# Print results

cat("Chi-squared test for multinomial distribution with unequal
probabilities:\n")

print(chisq.result)

# Interpret results

if (chisq.result$p.value > 0.05) {
  cat("p-value =", chisq.result$p.value,
      "\nWe fail to reject the null hypothesis of multinomial
fit.\n")
} else {
  cat("p-value =", chisq.result$p.value,
      "\nWe reject the null hypothesis of multinomial fit.\n")
}

```

APPENDIX D: R CODE FOR DATA AUGMENTATION

```
# Load libraries

library(copula)

library(MASS)

# Define the correlation

correlation <- 0.35

# Number of samples

n <- 1000

# Number of series

num_series <- 6

# Generate a multivariate normal distribution with the specified
correlation matrix

sigma <- matrix(correlation, num_series, num_series) + diag(1 -
correlation, num_series)

normals <- mvrnorm(n, mu = rep(0, num_series), Sigma = sigma)

# Transform the normal variables to a uniform using the CDF

uniforms <- pnorm(normals)

# Define the probabilities for the multinomial distributions

probs_list <- list(

  c(0.07, 0.43, 0.23, 0.27),

  c(0.39, 0.30, 0.15, 0.16),

  c(0.30, 0.26, 0.23, 0.21),

  c(0.38, 0.29, 0.20, 0.13),
```

```

    c(0.22, 0.35, 0.24, 0.19),
    c(0.36, 0.26, 0.16, 0.22))

# Ensure the number of probability vectors matches the number of
series

if (length(probs_list) != num_series) {

  stop("The length of probs_list must match the number of
series.")

}

# Function to map uniform variables to multinomial
uniform_to_multinomial <- function(u, probs) {

  return(findInterval(u, cumsum(probs), rightmost.closed = TRUE))

}

# Generate the series of multinomial random numbers
series_list <- lapply(1:num_series, function(i) {

  sapply(uniforms[, i], uniform_to_multinomial, probs =
probs_list[[i]])

}))

# Combine the series into a data frame
result <- as.data.frame(do.call(cbind, series_list))
colnames(result) <- paste0("series", 1:num_series)

# Save to a CSV file
write.csv(result, file = "GAD_synthetic.csv", row.names = TRUE)

# Verify the correlation
print(cor(result))

```

APPENDIX E: PYTHON CODE FOR ML MODELS

```
# Upload File

from google.colab import files

import pandas as pd

# Uploading the CSV file from computer

uploaded = files.upload()

filename = 'data.csv'

data = pd.read_csv(filename)

print(data)

# Separate features (questions) and target variable (evaluation)

X = data.drop('Target', axis=1)

y = data[' Target ' ]

##### RANDOM FOREST #####

import numpy as np

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, confusion_matrix, make_scorer

from sklearn.model_selection import cross_val_predict,
StratifiedKFold

import seaborn as sns

import matplotlib.pyplot as plt

import joblib
```

```

# Initialize the Random Forest model
model = RandomForestClassifier()

# Use Stratified K-Fold cross-validation
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Cross-validation predictions
y_pred = cross_val_predict(model, X, y, cv=skf)

model.fit(X, y)

# Calculate metrics
accuracy = accuracy_score(y, y_pred)
precision = precision_score(y, y_pred, average='weighted')
f1 = f1_score(y, y_pred, average='weighted')
recall = recall_score(y, y_pred, average='weighted')
conf_matrix = confusion_matrix(y, y_pred)

# Print the results
print(f"Random Forest Results:")
print(f"Accuracy: {accuracy:.2f}")
print(f"Precision: {precision:.2f}")
print(f"F1: {f1:.2f}")
print(f"Recall: {recall:.2f}")
print(f"Confusion Matrix:\n{conf_matrix}")

# Visualize the confusion matrix in a table form
conf_matrix_df = pd.DataFrame(conf_matrix, index=['Normal',
'Borderline', 'Clinical'], columns=['Predicted Normal',
'Predicted Borderline', 'Predicted Clinical'])

```



```

plt.figure(figsize=(10, 7))

sns.set(font_scale=1.2)

sns.heatmap(conf_matrix_df, annot=True, fmt='d', cmap='Blues',
            annot_kws={"size": 14})

# Adjusted annotation size

plt.ylabel('Actual', fontsize=12)

plt.xlabel('Predicted', fontsize=12)

plt.title('Confusion Matrix', fontsize=15)

plt.show()

# Save model

joblib.dump(model, f'{model}.joblib')

print(f"{model} trained and saved.")

print("\n")

##### DECISION TREE #####

import numpy as np

from sklearn.model_selection import cross_val_score,
cross_val_predict, StratifiedKFold

from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, confusion_matrix

from sklearn.tree import DecisionTreeClassifier

import seaborn as sns

import matplotlib.pyplot as plt

import joblib

```

```

# Use Stratified K-Fold cross-validation

skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Initialize the Decision Tree model

model = DecisionTreeClassifier()

# Perform cross-validation

y_pred = cross_val_predict(model, X, y, cv=skf)

model.fit(X, y)

accuracy = accuracy_score(y, y_pred)

precision = precision_score(y, y_pred, average='weighted')

f1 = f1_score(y, y_pred, average='weighted')

recall = recall_score(y, y_pred, average='weighted')

conf_matrix = confusion_matrix(y, y_pred)

# Print the results

print(f"Decision Tree Results:")

print(f"Accuracy: {accuracy:.2f}")

print(f"Precision: {precision:.2f}")

print(f"F1: {f1:.2f}")

print(f"Recall: {recall:.2f}")

print(f"Confusion Matrix:\n{conf_matrix}")

# Visualize the confusion matrix in a table form

conf_matrix_df = pd.DataFrame(conf_matrix, index=['Normal',
'Borderline', 'Clinical'], columns=['Predicted Normal',
'Predicted Borderline', 'Predicted Clinical'])

plt.figure(figsize=(10, 7))

```

```

sns.set(font_scale=1.2)

sns.heatmap(conf_matrix_df, annot=True, fmt='d', cmap='Blues',
            annot_kws={"size": 14}) # Adjusted annotation size

plt.ylabel('Actual', fontsize=12)

plt.xlabel('Predicted', fontsize=12)

plt.title('Confusion Matrix', fontsize=15)

plt.show()

# Save model

joblib.dump(model, f'{model}.joblib')

print(f"{model} trained and saved.")

print("\n")

##### SVM #####

import numpy as np

from sklearn.model_selection import cross_val_score,
cross_val_predict, StratifiedKFold

from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, confusion_matrix

from sklearn.svm import SVC

import seaborn as sns

import matplotlib.pyplot as plt

import joblib

# Initialize the SVM model

model = SVC()

```

```

# Use Stratified K-Fold cross-validation
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Cross-validation predictions
y_pred = cross_val_predict(model, X, y, cv=skf)

model.fit(X, y)

# Calculate metrics
accuracy = accuracy_score(y, y_pred)

precision = precision_score(y, y_pred, average='weighted')

f1 = f1_score(y, y_pred, average='weighted')

recall = recall_score(y, y_pred, average='weighted')

conf_matrix = confusion_matrix(y, y_pred)

# Print the results
print(f"SVM Results:")

print(f"Accuracy: {accuracy:.2f}")

print(f"Precision: {precision:.2f}")

print(f"F1: {f1:.2f}")

print(f"Recall: {recall:.2f}")

print(f"Confusion Matrix:\n{conf_matrix}")

# Visualize the confusion matrix in a table form

conf_matrix_df = pd.DataFrame(conf_matrix, index=['Normal',
'Borderline', 'Clinical'], columns=['Predicted Normal',
'Predicted Borderline', 'Predicted Clinical'])

plt.figure(figsize=(10, 7))

sns.set(font_scale=1.2)

```

```

sns.heatmap(conf_matrix_df, annot=True, fmt='d', cmap='Blues',
annot_kws={"size": 14}) # Adjusted annotation size

plt.ylabel('Actual', fontsize=12)

plt.xlabel('Predicted', fontsize=12)

plt.title('Confusion Matrix', fontsize=15)

plt.show()

# Save model

joblib.dump(model, f'{model}.joblib')

print(f"{model} trained and saved.")

print("\n")

##### LOGISTIC REGRESSION #####

import numpy as np

from sklearn.model_selection import cross_val_score,
cross_val_predict, StratifiedKFold

from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, confusion_matrix

from sklearn.linear_model import LogisticRegression

import seaborn as sns

import matplotlib.pyplot as plt

import joblib

# Use Stratified K-Fold cross-validation

skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Initialize the Logistic Regression model

```

```

model = LogisticRegression(max_iter=1000)

# Cross-validation predictions
y_pred = cross_val_predict(model, X, y, cv=skf)

model.fit(X, y)

# Calculate metrics
accuracy = accuracy_score(y, y_pred)

precision = precision_score(y, y_pred, average='weighted')

f1 = f1_score(y, y_pred, average='weighted')

recall = recall_score(y, y_pred, average='weighted')

conf_matrix = confusion_matrix(y, y_pred)

# Print the results

print(f"Logistic Regression Results:")

print(f"Accuracy: {accuracy:.2f}")

print(f"Precision: {precision:.2f}")

print(f"F1: {f1:.2f}")

print(f"Recall: {recall:.2f}")

print(f"Confusion Matrix:\n{conf_matrix}")

# Visualize the confusion matrix in a table form

conf_matrix_df = pd.DataFrame(conf_matrix, index=['Normal',
'Borderline', 'Clinical'], columns=['Predicted Normal',
'Predicted Borderline', 'Predicted Clinical'])

plt.figure(figsize=(10, 7))

sns.set(font_scale=1.2)

```

```

sns.heatmap(conf_matrix_df, annot=True, fmt='d', cmap='Blues',
annot_kws={"size": 14}) # Adjusted annotation size

plt.ylabel('Actual', fontsize=12)

plt.xlabel('Predicted', fontsize=12)

plt.title('Confusion Matrix', fontsize=15)

plt.show()

# Save model

joblib.dump(model, f'{model}.joblib')

print(f"{model} trained and saved.")

print("\n")

##### NAIVE BAYES #####

import numpy as np

from sklearn.model_selection import cross_val_score,
cross_val_predict, StratifiedKFold

from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, confusion_matrix

from sklearn.naive_bayes import GaussianNB

import seaborn as sns

import matplotlib.pyplot as plt

import joblib

# Use Stratified K-Fold cross-validation

skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Initialize the Naive Bayes model

```

```

model = GaussianNB()

# Perform cross-validation

y_pred = cross_val_predict(model, X, y, cv=skf)

model.fit(X, y)

accuracy = accuracy_score(y, y_pred)

precision = precision_score(y, y_pred, average='weighted')

f1 = f1_score(y, y_pred, average='weighted')

recall = recall_score(y, y_pred, average='weighted')

conf_matrix = confusion_matrix(y, y_pred)

# Print the results

print(f"Naive Bayes Results:")

print(f"Accuracy: {accuracy:.2f}")

print(f"Precision: {precision:.2f}")

print(f"F1: {f1:.2f}")

print(f"Recall: {recall:.2f}")

print(f"Confusion Matrix:\n{conf_matrix}")

# Visualize the confusion matrix in a table form

conf_matrix_df = pd.DataFrame(conf_matrix, index=['Normal',
'Borderline', 'Clinical'], columns=['Predicted Normal',
'Predicted Borderline', 'Predicted Clinical'])

plt.figure(figsize=(10, 7))

sns.set(font_scale=1.2)

# Adjust the font scale for better readability

```



```

sns.heatmap(conf_matrix_df, annot=True, fmt='d', cmap='Blues',
annot_kws={"size": 14}) # Adjusted annotation size

plt.ylabel('Actual', fontsize=12)

plt.xlabel('Predicted', fontsize=12)

plt.title('Confusion Matrix', fontsize=15)

plt.show()

# Save model

joblib.dump(model, f'{model}.joblib')

print(f"{model} trained and saved.")

print("\n")

##### KNN #####

import numpy as np

from sklearn.model_selection import cross_val_score,
cross_val_predict, StratifiedKFold

from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, confusion_matrix

from sklearn.neighbors import KNeighborsClassifier

import seaborn as sns

import matplotlib.pyplot as plt

import joblib

# Use Stratified K-Fold cross-validation

skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Initialize the KNN model

```

```

model = KNeighborsClassifier()

# Perform cross-validation

y_pred = cross_val_predict(model, X, y, cv=skf)

model.fit(X, y)

accuracy = accuracy_score(y, y_pred)

precision = precision_score(y, y_pred, average='weighted')

f1 = f1_score(y, y_pred, average='weighted')

recall = recall_score(y, y_pred, average='weighted')

conf_matrix = confusion_matrix(y, y_pred)

# Print the results

print(f"KNN Results:")

print(f"Accuracy: {accuracy:.2f}")

print(f"Precision: {precision:.2f}")

print(f"F1: {f1:.2f}")

print(f"Recall: {recall:.2f}")

print(f"Confusion Matrix:\n{conf_matrix}")

# Visualize the confusion matrix in a table form

conf_matrix_df = pd.DataFrame(conf_matrix, index=['Normal',
'Borderline', 'Clinical'], columns=['Predicted Normal',
'Predicted Borderline', 'Predicted Clinical'])

plt.figure(figsize=(10, 7))

sns.set(font_scale=1.2)

sns.heatmap(conf_matrix_df, annot=True, fmt='d', cmap='Blues',
annot_kws={"size": 14})

```

```

# Adjusted annotation size
plt.ylabel('Actual', fontsize=12)
plt.xlabel('Predicted', fontsize=12)
plt.title('Confusion Matrix', fontsize=15)
plt.show()

# Save model
joblib.dump(model, f'{model}.joblib')
print(f"{model} trained and saved.")
print("\n")

##### EXTERNAL VALIDATION #####

from google.colab import files

import pandas as pd

from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, confusion_matrix

import joblib

# Load data

uploaded = files.upload()

filename = 'test data.csv'

data = pd.read_csv(filename)

print(data)

X_test = data.drop("Total_Elevation", axis=1)

y_test = data["Total_Elevation"]

```

```

# List of models to test

models = ['RandomForestClassifier()', 'DecisionTreeClassifier()',
'SVC()', 'LogisticRegression(max_iter=1000)', 'GaussianNB()',
'KNeighborsClassifier()']

for model_name in models:

    # Load model

    model = joblib.load(f'{model_name}.joblib')

    # Predict on test data

    y_pred = model.predict(X_test)

    # Calculate metrics

    accuracy = accuracy_score(y_test, y_pred)

    precision = precision_score(y_test, y_pred, average='weighted')

    recall = recall_score(y_test, y_pred, average='weighted')

    f1 = f1_score(y_test, y_pred, average='weighted')

    confusion = confusion_matrix(y_test, y_pred)

    # Output metrics

    print(f"Model: {model_name}")

    print(f"Accuracy: {accuracy:.2f}")

    print(f"Precision: {precision:.2f}")

    print(f"Recall: {recall:.2f}")

    print(f"F1 Score: {f1:.2f}")

    print(f"Confusion Matrix:\n{confusion}")

    print("\n")

```