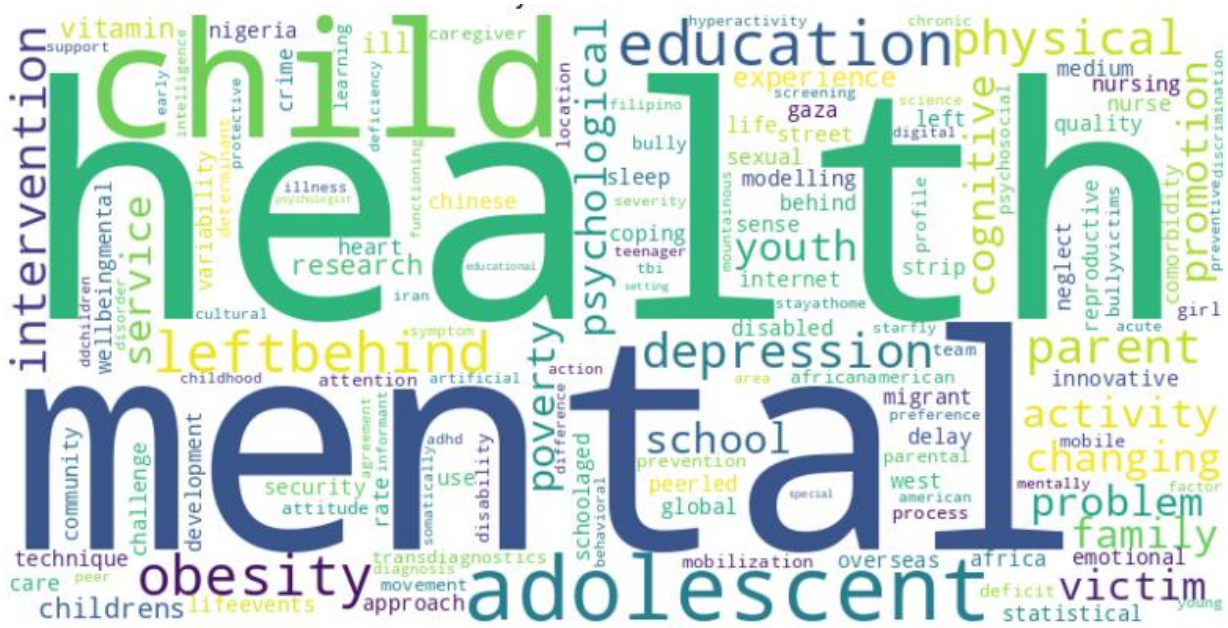
Data Analytics Project

Children Mental Health Analysis

404 Not Found



Group Members:

Dai-Chi Wang, Zhiyi Wang, Yiwen Ge, Ning Ding, Jialu Li

1. Background and context to the problem statement

In today's digital age, problematic internet use among children and adolescents is increasingly linked to mental health issues like depression and anxiety. Research shows that teens spend an average of 7 hours and 22 minutes on screens daily, while younger children aged 8–12 spend around 4 hours and 44 minutes. In the U.S., mental health concerns affect 11.5% of youth —that's over 2.7 million young individuals. These challenges are closely connected, as too much internet use often leads to mental health problems. Our project focuses on this phenomenon, aiming to help identify factors that contribute to children's mental health issues, enabling timely interventions and promoting healthier digital habits.

To address this issue, a two-part approach was implemented. First, relevant research articles were scraped for text mining analysis. Second, a Kaggle dataset containing physical activity and fitness data was analyzed using machine learning techniques. By comparing insights from both methods, this study aims to provide a comprehensive understanding of the topic and inform future actions.

2. Data Acquisition

2.1 Web Scraping

The web scraping retrieved relevant research articles to explore the origins of children's mental health issues. By extracting and filtering data from an open-access library, the study identified 36 articles from an initial pool of 500, focusing on topics directly related to mental health and children. This curated dataset included essential information such as keywords and abstracts, ensuring that the analysis remained targeted and aligned with the research objectives.

2.2 Kaggle

The data is from child mind institute. This dataset proposes using physical fitness indicators (such as posture, diet, and physical activity) as proxies to identify excessive technology use. Our objective is to analyze the factors within the dataset that influence the Severity Impairment Index (SII). By leveraging physical fitness indicators such as posture, diet, and physical activity, alongside internet usage behavior data, we aim to uncover key relationships and patterns. The ultimate goal is to develop a predictive model that identifies early signs of mental health among children and young adults, enabling timely interventions to promote healthier habits and overall well-being.

3. Data Analysis

3.1 Text Mining

We first performed the text analysis for the articles that we extracted. We cleaned and preprocessed both Keywords and Abstracts of those articles to standardize and dropped all unnecessary text such as stopwords, punctuations and non-English words.

(1) Keyword Analysis

The first text analysis generated the top 20 most common keywords from the Keywords and plotted a bar chart to visualize the frequency as in 3.1.1. We found other than the obvious ones which are exactly of focus such as mental, health and child, education, leftbehind and obesity are also quite high, indicating the physical condition and feelings in school are quite highly correlated with mental health problems. Furthermore, to oversee all high frequency words, we also drawed a word cloud 3.1.2, and found words like parent, family, poverty, internet are all possible contributors as well.

(2) Abstract Analysis

Then, to analyze the abstract part, which is the main content of each article, we performed a key term frequency analysis as well. To make the result more meaningful and accurate, we grouped the term with similar meanings together by similarity threshold 0.7. We assumed five possible factors based on keywords analysis and our understanding of the kaggle dataset: Internet, Sleep, Sex, Body, Exercise, and Sex. We also manually added some specific term to each group such as insomnia to sleep. We then performed a bar chart 3.1.3 for those grouped terms and found sex is the highest, and much higher than other categories among all. The Internet is the second highest, and next is exercise.

(3) Collocation Analysis

The collocation analysis identifies significant bigrams 3.1.4 in the dataset, with '**mental health**' scoring the highest (590.80), highlighting a strong focus on this topic, as we expected. Other notable bigrams, such as '**physical activity**' (90.46), '**sense of security**' (76.27), and '**sexual reproductive**' (57.80), point to themes related to health, security, and societal issues. Phrases like '**mobile team**' and '**statistically significant**' suggest discussions on mobility and statistical analysis. This indicates that physical activity and psychological feeling are highly correlated with teenagers' mental health problems.

(4) LDA Analysis

The LDA topic modeling 3.1.5 identifies five key themes from the dataset. Topic 1 focuses on adolescent mental health and interventions, while Topic 2 highlights mental health challenges like discrimination and sleep issues. Topic 3 centers on child and adolescent behavior and family dynamics. Topic 4 explores parental influence, physical health, and sleep, and Topic 5 addresses health conditions, including sexual health. Most abstracts are linked to Topic 3, reflecting its prominence in the dataset. These insights highlight sleep and discrimination problems are important factors.

3.2 Exploratory Data Analysis

(1) Distribution of Target Variable

The distribution of the Severity Impairment Index (SII) shows a predominance of participants categorized as having "None" impairment, followed by "Mild" and "Moderate" levels. Severe cases are rare. A significant portion of the data is marked as missing, as seen in 3.2.1. This imbalance should be considered during model development. The PCIAT_Total Score (Problematic Internet Use) exhibits a right-skewed distribution (see 3.2.2), indicating that most participants have lower scores.

(2) Gender and Age Distribution

As seen in 3.2.3, the dataset contains more male participants than female participants. Additionally, the age distribution (3.2.4) suggests that the sample is skewed toward younger individuals, with the majority falling between ages 7 and 15.

(3) Variables Distribution

Participants' internet usage patterns vary widely, with most reporting less than 1 hour per day (3.2.5). A smaller portion of the sample reports more extensive usage, which might indicate problematic behavior.

Physical metrics such as BMI, weight, height, blood pressure, and heart rate exhibit variations with some outliers present (3.2.6). These outliers may require further investigation or preprocessing.

Sleep disturbance scores (3.2.7) indicate that most participants fall within a normal range, but there are individuals with significantly high scores.

4. Machine Learning

4.1 Preprocessing

(1) Correlation Analysis

The heatmap (4.1.1) highlights the correlations between continuous variables and the target variable (SII). Variables such as body metrics, physical activity, and sleep indicators exhibit significant relationships with SII, making them critical for modeling. We deleted the unrelated columns like the seasons variables.

(2) Handling Missing Values

Categorical Columns: Mode imputation was used with the most frequent category.

Numerical Columns: Mean imputation was applied in continuous variables.

Target Variable: Rows missing SII were dropped to maintain the integrity of the target data.

(3) Train-Test Splitting

The dataset was split into training and testing sets using a Stratified Split Method. This approach ensured that the distribution of the target variable (SII) was consistent across both subsets, mitigating risks related to data imbalance during model training and evaluation.

(4) Dimensionality Reduction

To manage the high dimensionality of the dataset, Principal Component Analysis (PCA) was performed, resulting in 17 principal components:

PC1: Hydration; **PC2**: Body Metrics; **PC3**: Fitness; **PC4**: Sleep; **PC5**: Sleep Endurance; **PC6**: Fat Mass; **PC7**: Fat-Free Mass; **PC8**: Grip Strength; **PC9**: Curl Up; **PC10**: Physical Activity_by Child; **PC11**: Physical Activity_by Parents; **PC12**: Fitness Duration; **PC13**: Fitness Endurance; **PC14**: Cardiovascular; **PC15**: Waist Circumference; **PC16**: Trunk Strength; **PC17**: Push-up

4.2 Machine Learning Models

We applied 4 supervised models and the details will be discussed in the following sections.

(1) Linear Regression Model

The model's current predictive performance appears limited, as indicated by the R-squared scores (Training: 0.2369, Validation: 0.1021) and Mean Squared Error (MSE) values (Training: 0.4507, Validation: 0.5465). These results suggest that there may not be a strong linear relationship between the features and the target variable.

(2) Random Forest

The model demonstrates moderate performance, with the best parameters being a depth of 3 and a minimum of 1 sample per leaf. The accuracy results show a training accuracy of 0.5868 and a testing accuracy of 0.6277, indicating that the model generalizes well to unseen data as the testing accuracy slightly exceeds the training accuracy.

(3) XGBoost

The XGBoost model demonstrates relatively strong performance with the best parameters being 100 estimators, a depth of 1, and a learning rate of 1. The accuracy results indicate a training accuracy of 0.7061 and a testing accuracy of 0.6259. While the model performs better on the training data, the gap suggests a potential overfitting issue.

(4) Neural Network

Neural network achieves 59.87% accuracy, which is comparable to the performance of XGBoost and Random Forest. The model was trained using the selected hyperparameters: the 'tanh' activation function, adaptive learning rate, 30 hidden units, 2000 iterations, and 'adam' solver.

4.3 Important Features Comparison

We identified the top ten most important features for each of the four models. In the analysis of various models, the Linear Regression model revealed that "Sit & Reach fitness zone" had the highest score, but the R^2 score of approximately 20% suggests that most features have a minimal impact on the target variable. For the Random Forest model, "Hydration" (PC1) emerged as the most important feature, with other physical factors such as "PC2," "PC8," and "PC5," along with "Internet Use," ranking among the top five. In the XGBoost model, "Internet Use" was identified as the most significant feature, followed by "Hydration" (PC1) and "Sex of participant" in second and third positions, respectively. Similarly, in the Neural Networks model, "Internet Use" was the most important feature, with "Sleep" (PC4) and "Sex of participant" taking the second and third spots.

To conclude, based on the analysis of the four models, it is clear that "Internet Use," "Hydration," "Sex," and "Sleep" are consistently identified as the most important features across different modeling approaches.

5. Conclusion

Our research investigates the relationship between problematic internet use and mental health issues in children and adolescents through a dual approach: text-mining analysis of research articles and machine learning on behavioral and physical activity data.

Our text-mining analysis highlights the broader contextual factors influencing youth mental health. Keywords like "internet," "sleep," "family," and "physical health" appear frequently. The finding illustrates the multifaceted nature of mental health challenges, where behavioral habits, family dynamics, and societal factors interplay.

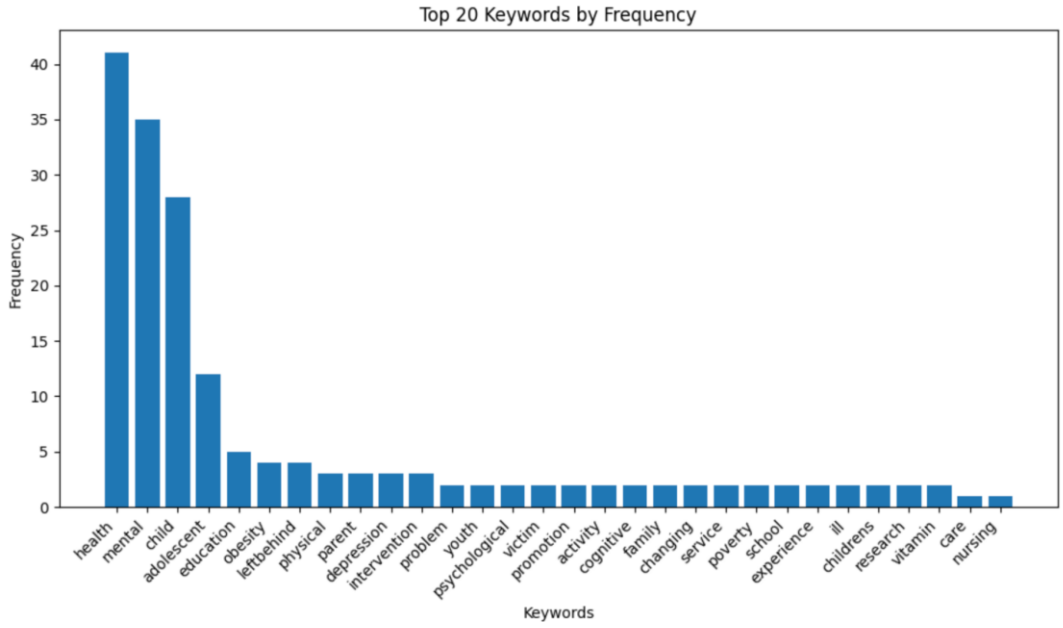
In comparison, Machine learning models quantitatively confirm the significance of internet usage, sleep, hydration, and sex in predicting mental health outcomes. In the better-performing XGBoost model, "Internet Use," "Hydration," and "Sex" are identified as significant features, with "Internet Use" ranking as the most critical predictor. The prominence of "Sex" includes sexual harassment, which, although infrequent, can have a substantial effect on mental health.

Based on our results, parents play a pivotal role in safeguarding children's mental health in the digital age. To prevent potential issues, parents should encourage their children to set clear guidelines for internet use, ensure regular exercise and establish consistent sleep routines. In cases where severe incidents, such as exposure to sexual harm, have already occurred, parents should seek professional psychological counseling and support their child’s recovery as soon as possible. By combining preventative strategies with timely intervention, parents can provide a secure and nurturing environment that promotes their children's resilience and mental health.

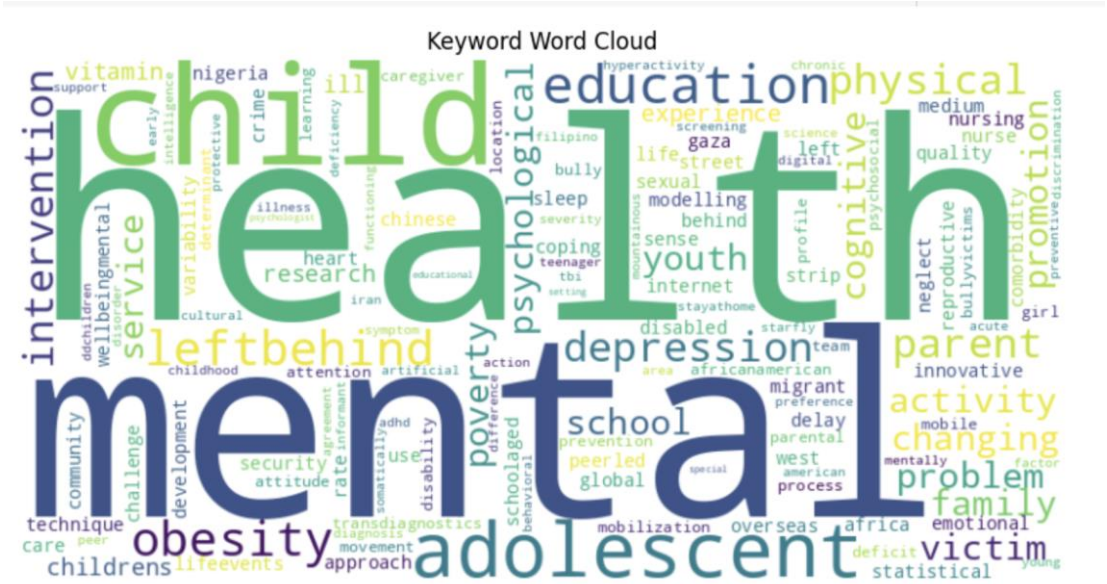
Appendix

Link	Keywords	Abstract
https://www.	Child Care, C	This study was carried out in order to investigate the professionals,Âõ feelings concerning the care in the Center for Children and Youth Psychosocial Care, as well as to evidence the advantages and difficulties of this work. This was a q
https://www.	Disabled Chi	The aim of the study was to investigate the prevalence of mental health problems and quality of life among Palestinian disable children. The sample consisted of 391 disable Palestinian children in the Gaza Strip which was selected ran
https://www.	Internet, Phy	The purpose of the research was to study the behavior of children and youth related to using the internet. The sample consisted of 1,584 children and youths in Bangkok. The Canonical Correlation Analysis statistics method was applied
https://www.	Psychologica	Background:In Asia, especially in China and ASEAN,ithas been an obvious trend that the population in general and parents in particular leave the countryside to bigger cities or more developed countries for employment opportunities.
https://www.	Street, Child	Background:Street life is a common sight among children in Africa including Nigeria. Distressed, hungry and poorly groomed children can be seen roaming the streets in search for means of survival from well-wishers and passersby. Pov
https://www.	Mental Healt	Every woman, man, youth and child has the human right to the highest attainablestandard of physical and mental health, without discrimination of any kind. This isenshrined in our Indian Constitution and the Universal Declaration on H
https://www.	Obesity, Chil	A number of studies have reported significant associations between obesity and poor psychological wellbeing in children but findings have been inconsistent. Methods: This study utilised data from 3,898 children aged 5-16 years obtair
https://www.	Overseas Ch	This paper surveys 715 children who have been left behind in China by their √©migr√© parents. This research was conducted using the MMHI-60, ALSEC and S-I questionnaire and the findings are as follows: 1) the general mental healt
https://www.	Nurse Educa	Background:The challenge for nursing training highlights the priorities for knowing and attending current educational tendencies on adolescents and the universitycurriculum, required for a nurse professional to develop skills to reach
https://www.	Sleep Proble	This study explored characteristics of sleep and other presenting problems in children and adolescents seeking mental health services within an outpatient clinic. Primary caregivers seeking outpatient mental health services for their
https://www.	Adolescent H	Background:India being the country with largest adolescent population in the world, needs a special focus on the health services as the status of an adolescent determines the health status in his/her adulthood. Some of the major issu
https://www.	Health Promo	Many health promotion programs have been implemented to prevent obesity and mental health problems among school-aged children. However, only a few programs included both physical and psychological measures to assess the ef
https://www.	Child and Ad	Child and adolescent mental health is an essential component of overall health and its importance is gaining increased recognition. Current events have heightened an interest in the mental health of youth. Unfortunately, too often th

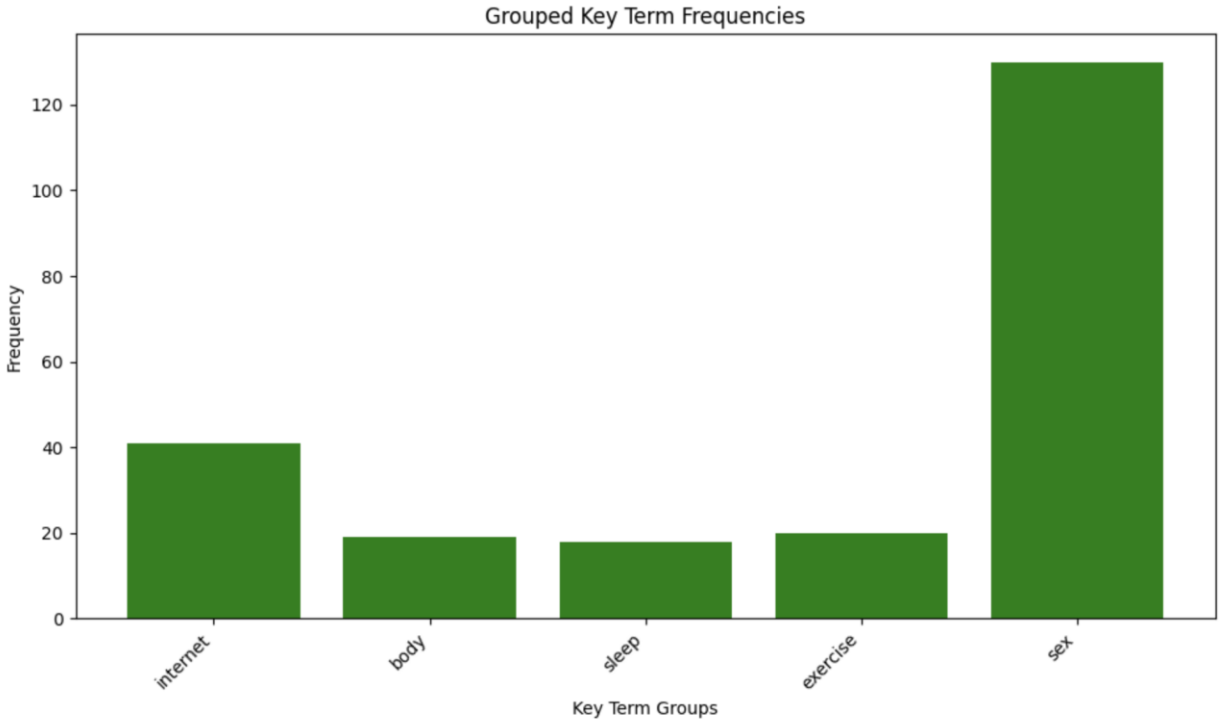
2.1.1 Results of Web Scraping



3.1.1 Top 20 Keywords by Frequency



3.1.2 Keywords Word Cloud



3.1.3 Grouped Key Term Frequencies

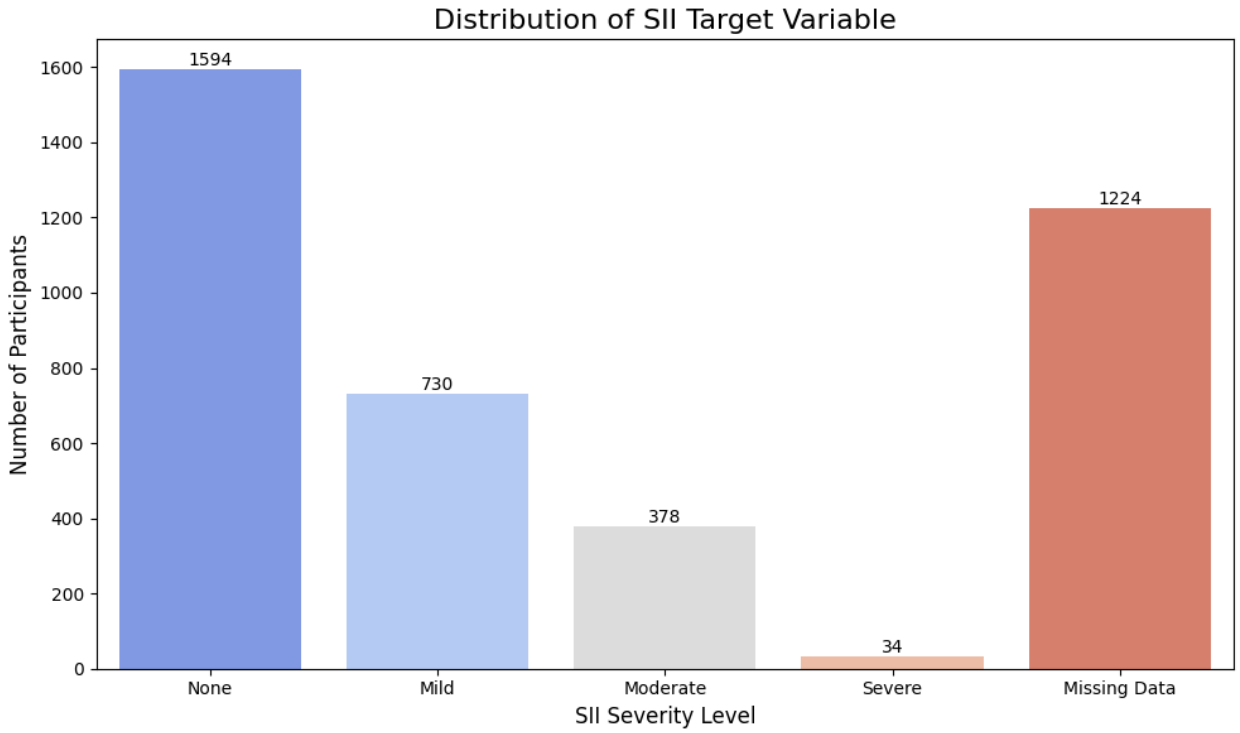
```
Top 10 Collocations (Bigrams):
('mental', 'health'): 590.798871609772
('physical', 'activity'): 90.4551568414237
('sense', 'security'): 76.27379708713335
('mobile', 'team'): 74.18090573356012
('life', 'event'): 62.43276405748617
('sexual', 'reproductive'): 57.80122088649254
('mobile', 'office'): 55.71204109286172
('statistically', 'significant'): 55.676682108475006
('symptom', 'severity'): 54.37310023237684
('rural', 'area'): 52.79772645244116
```

3.1.4 Top 10 Collocations

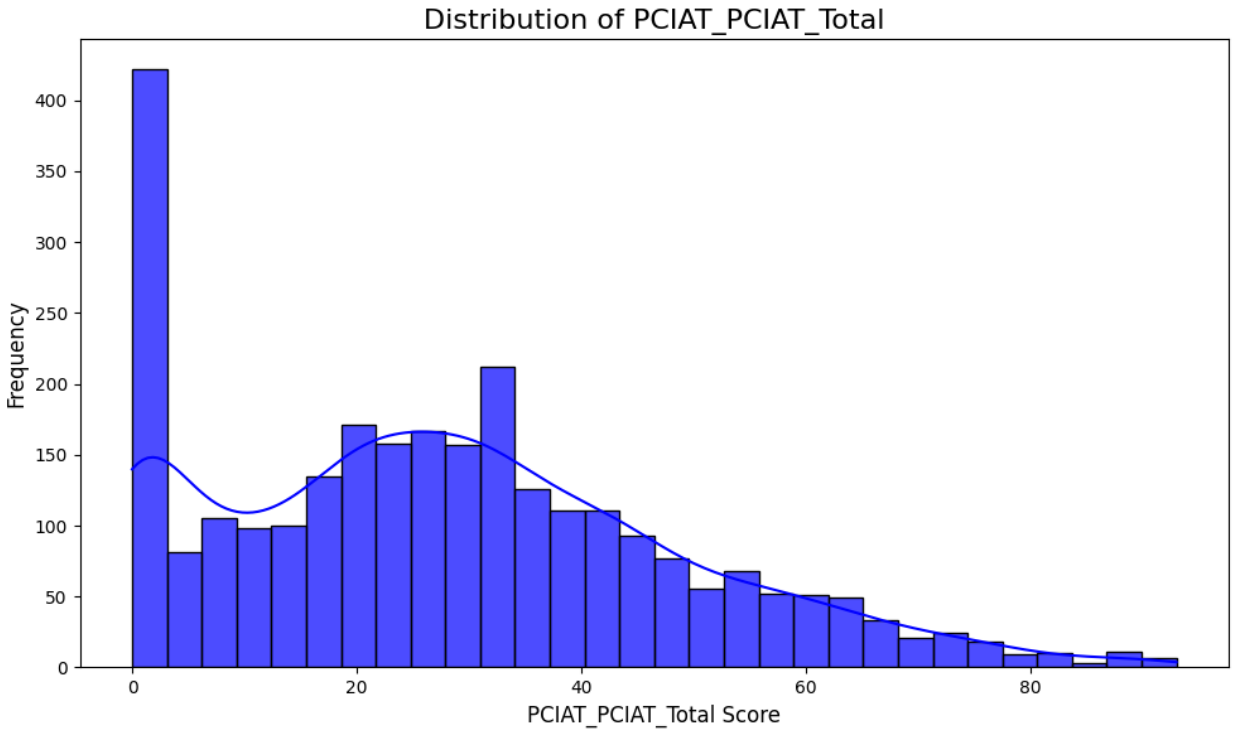
```
LDA Topics with Key Terms:
Topic 1: ['outcome', 'girl', 'adolescent', 'problem', 'diagnosis', 'intervention', 'group', 'child', 'mental', 'health']
Topic 2: ['discrimination', 'mental', 'impact', 'level', 'parental', 'sleep', 'association', 'study', 'problem', 'symptom']
Topic 3: ['functioning', 'mean', 'study', 'parent', 'childrens', 'adolescent', 'family', 'mental', 'child', 'health']
Topic 4: ['adolescent', 'parent', 'physical', 'effect', 'sleep', 'program', 'study', 'mental', 'health', 'child']
Topic 5: ['condition', 'disease', 'participant', 'intervention', 'sexual', 'issue', 'study', 'adolescent', 'mental', 'health']

Linked Topics to Abstracts:
                                        Abstract  dominant_topic
0  study carried order investigate professional f...               3
1  aim study investigate prevalence mental health...               2
2  purpose research study behavior child youth re...               3
3  backgroundin asia especially china aseanithas ...               3
4  backgroundstreet life common sight among child...               3
```

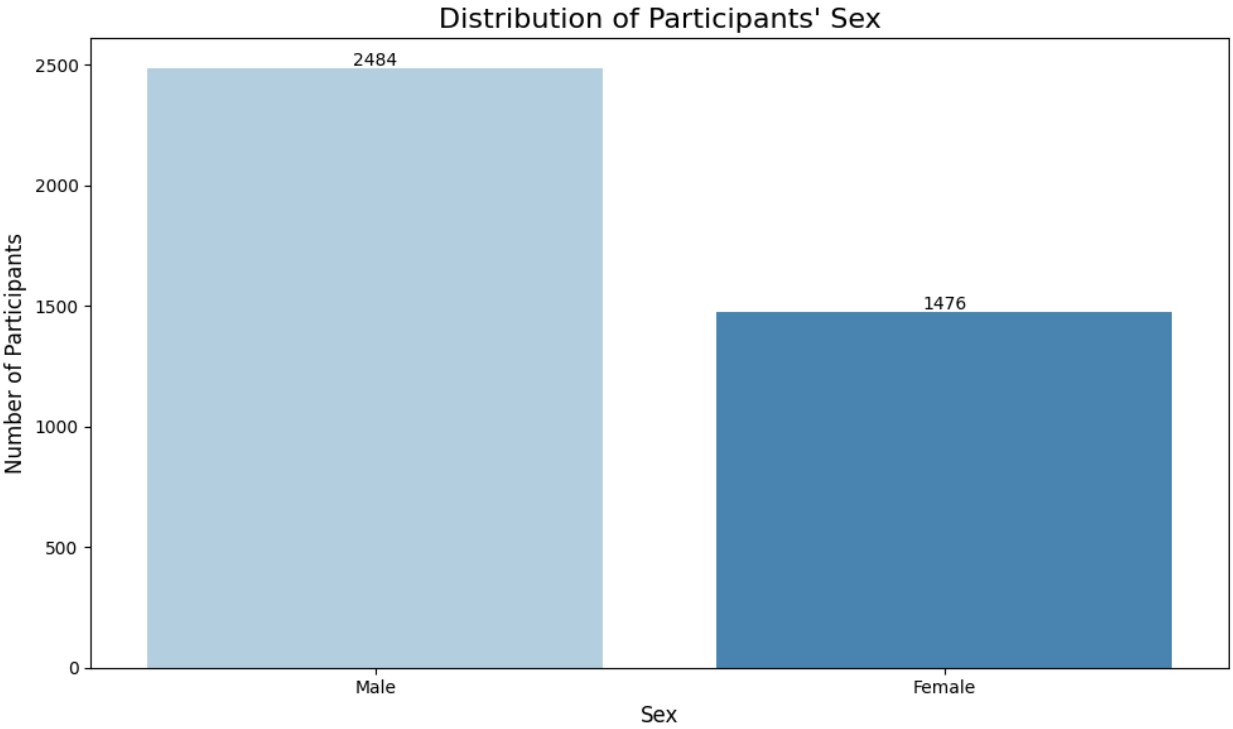
3.1.5 LDA analysis



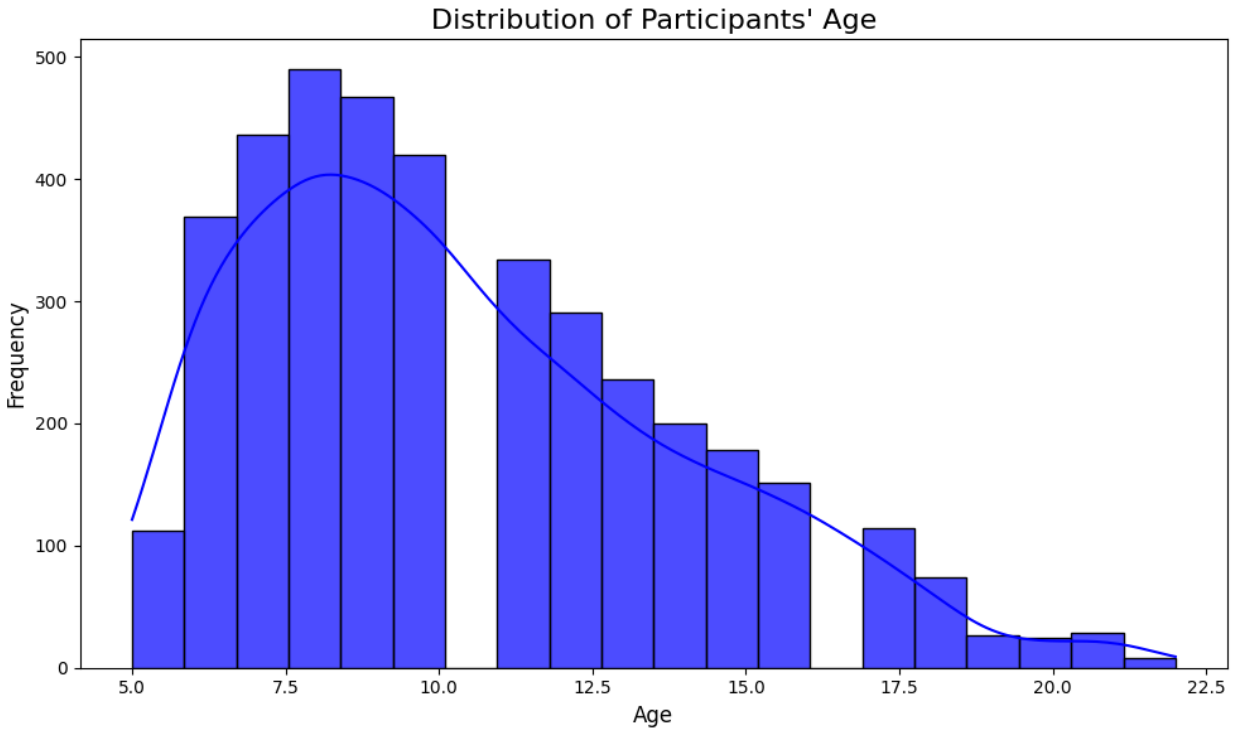
3.2.1 Distribution of SII Target Variable



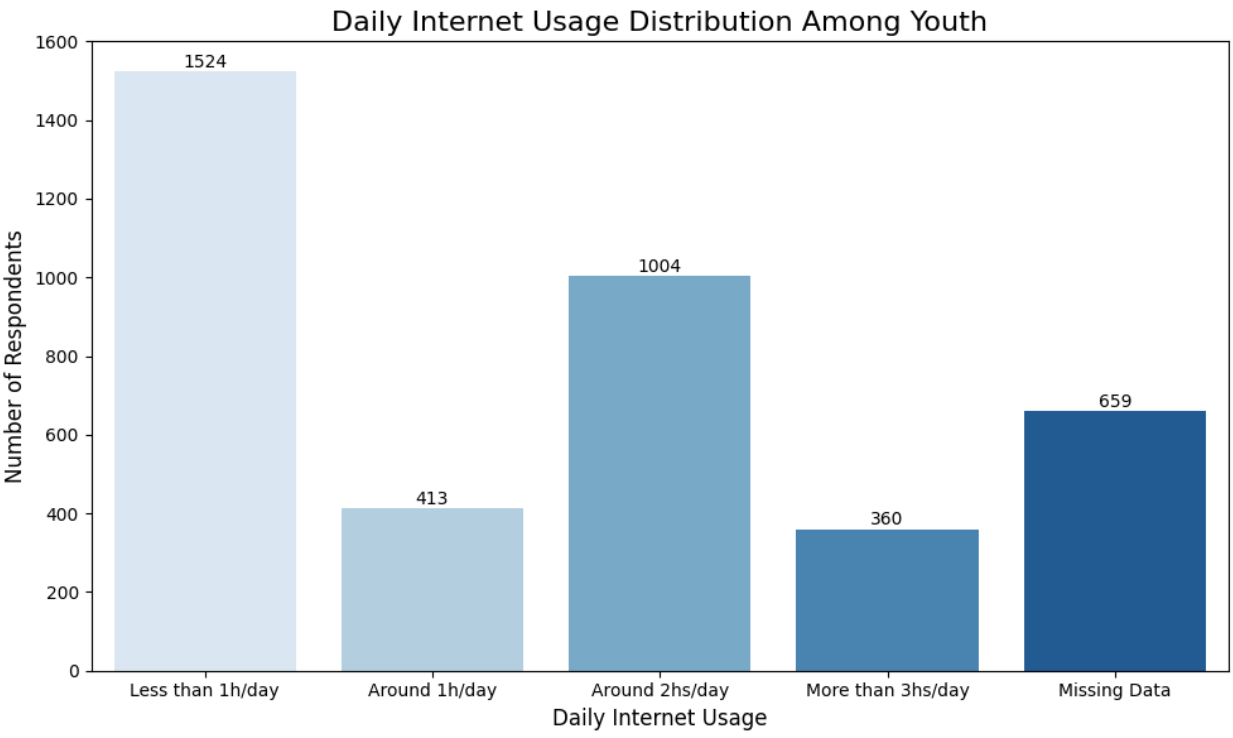
3.2.2 Distribution of PCIAT_PCIAT_Total



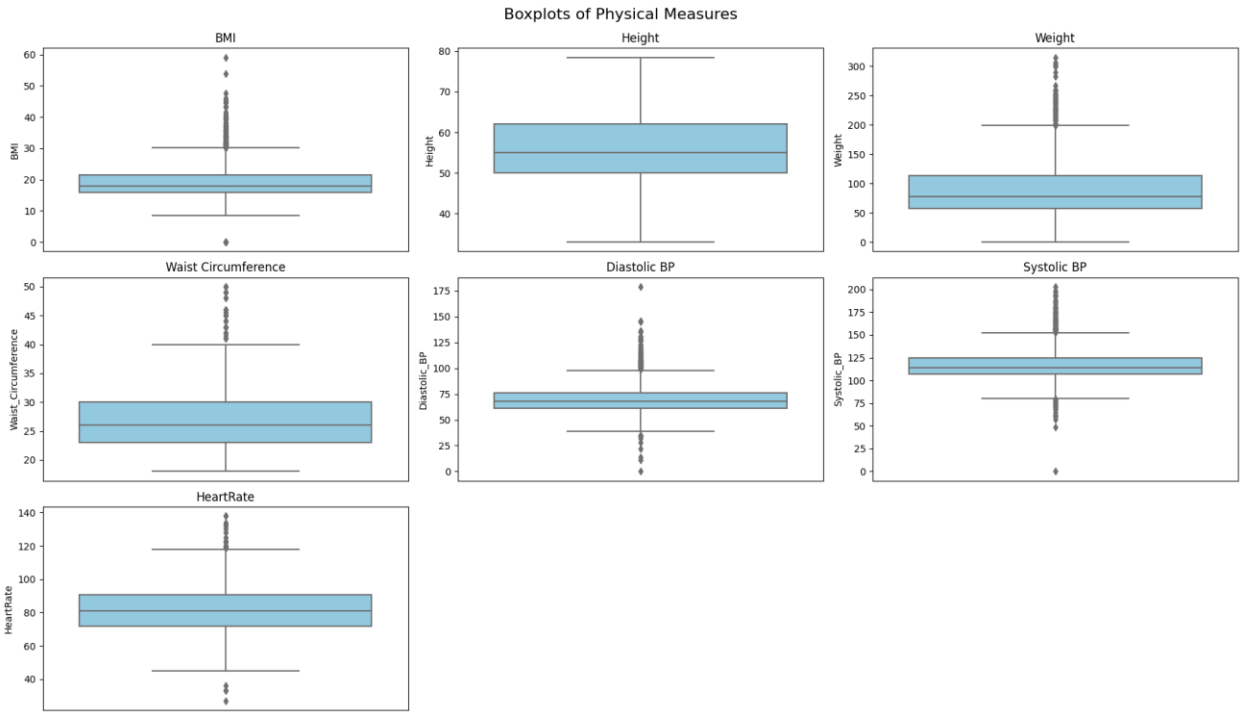
3.2.3 Gender Distribution



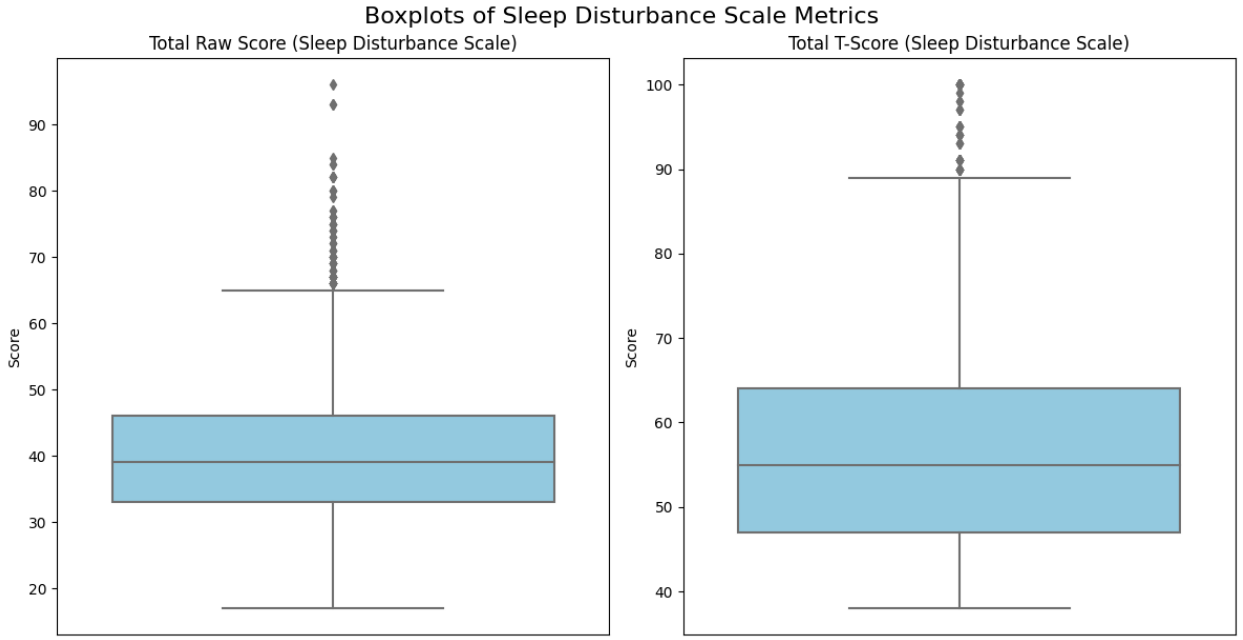
3.2.4 Age Distribution



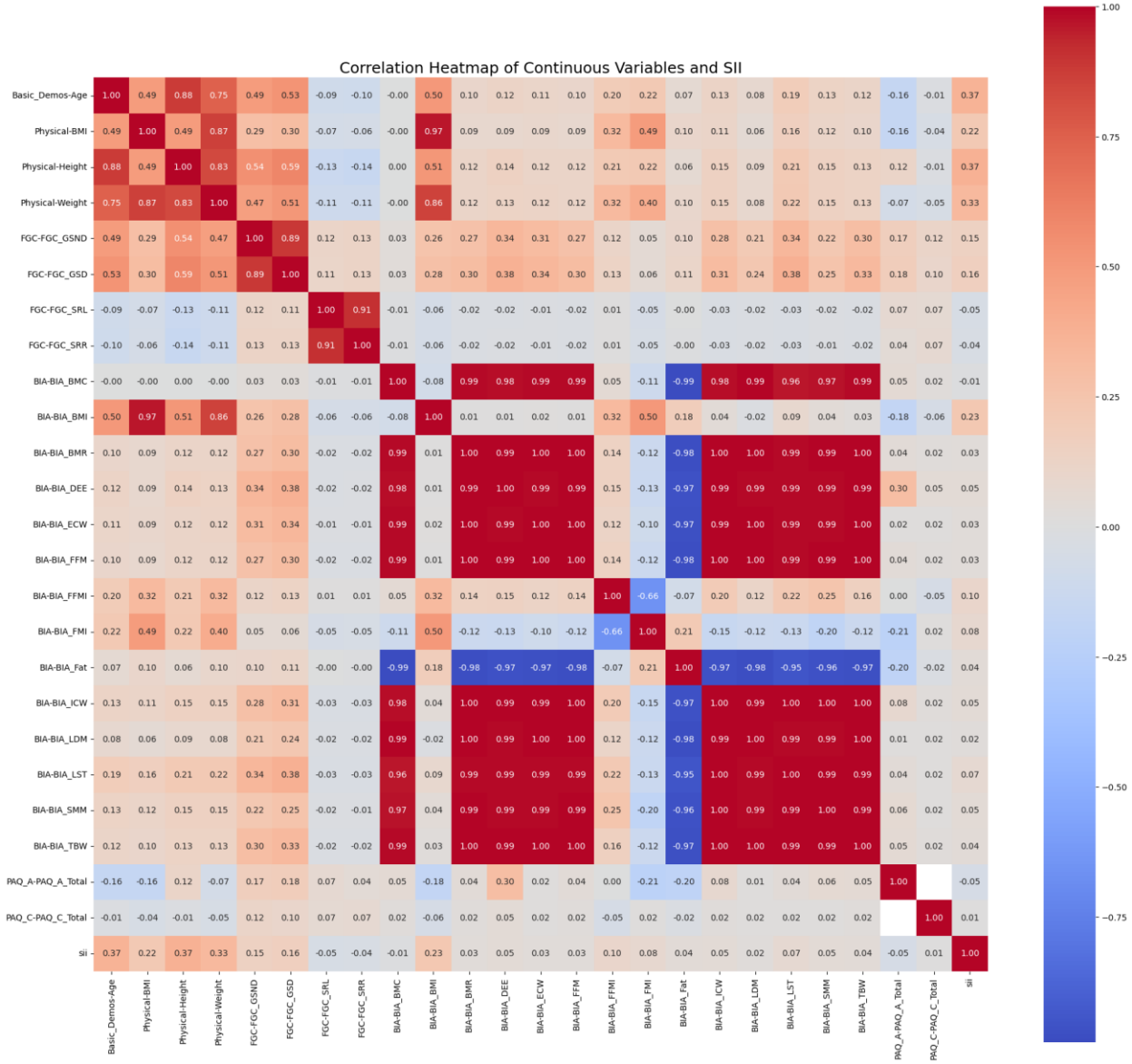
3.2.5 Distribution of Internet Usage



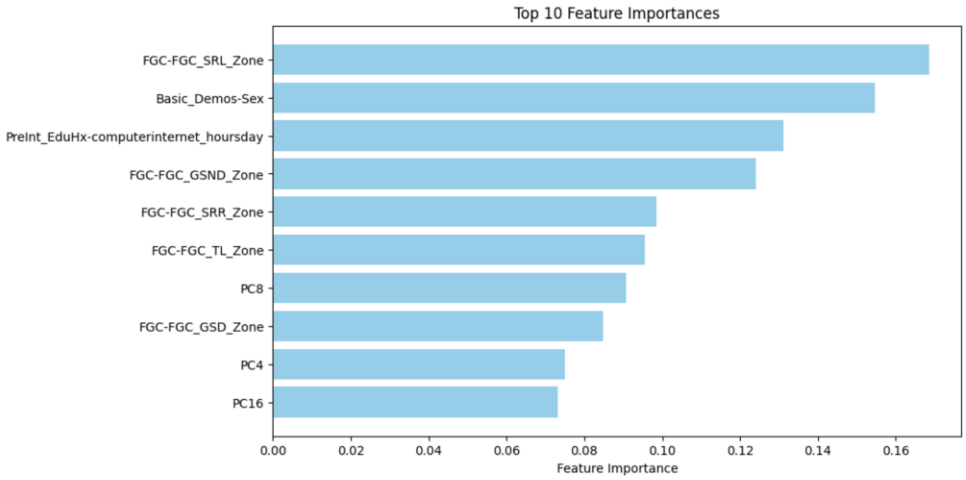
3.2.6 Distribution of Physical Measures



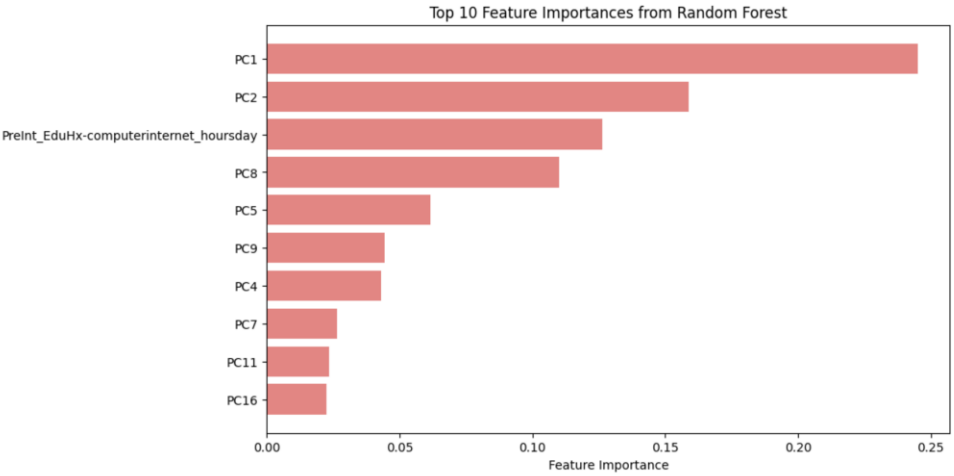
3.2.7 Distribution of Sleep



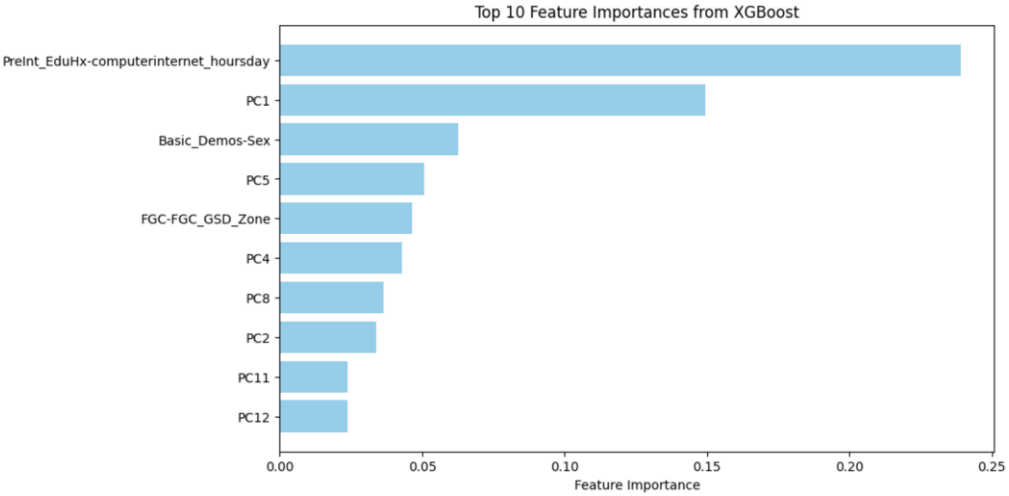
4.1.1 Heatmap of Variables



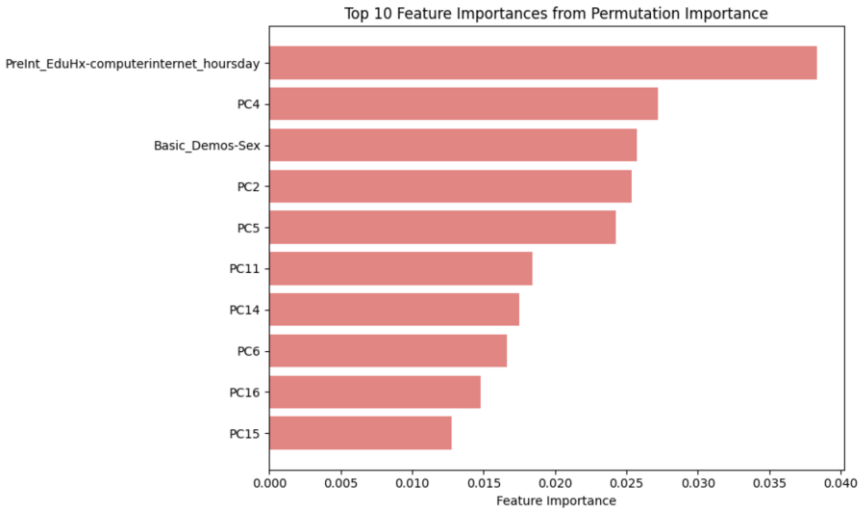
4.3.1 Linear Regression- Top 10 Feature Importance



4.3.2 Random Forest-Top 10 Feature Importance



4.3.3 XGBoost-Top 10 Feature Importance



4.3.4 Neural Network-Top 10 Feature Importance