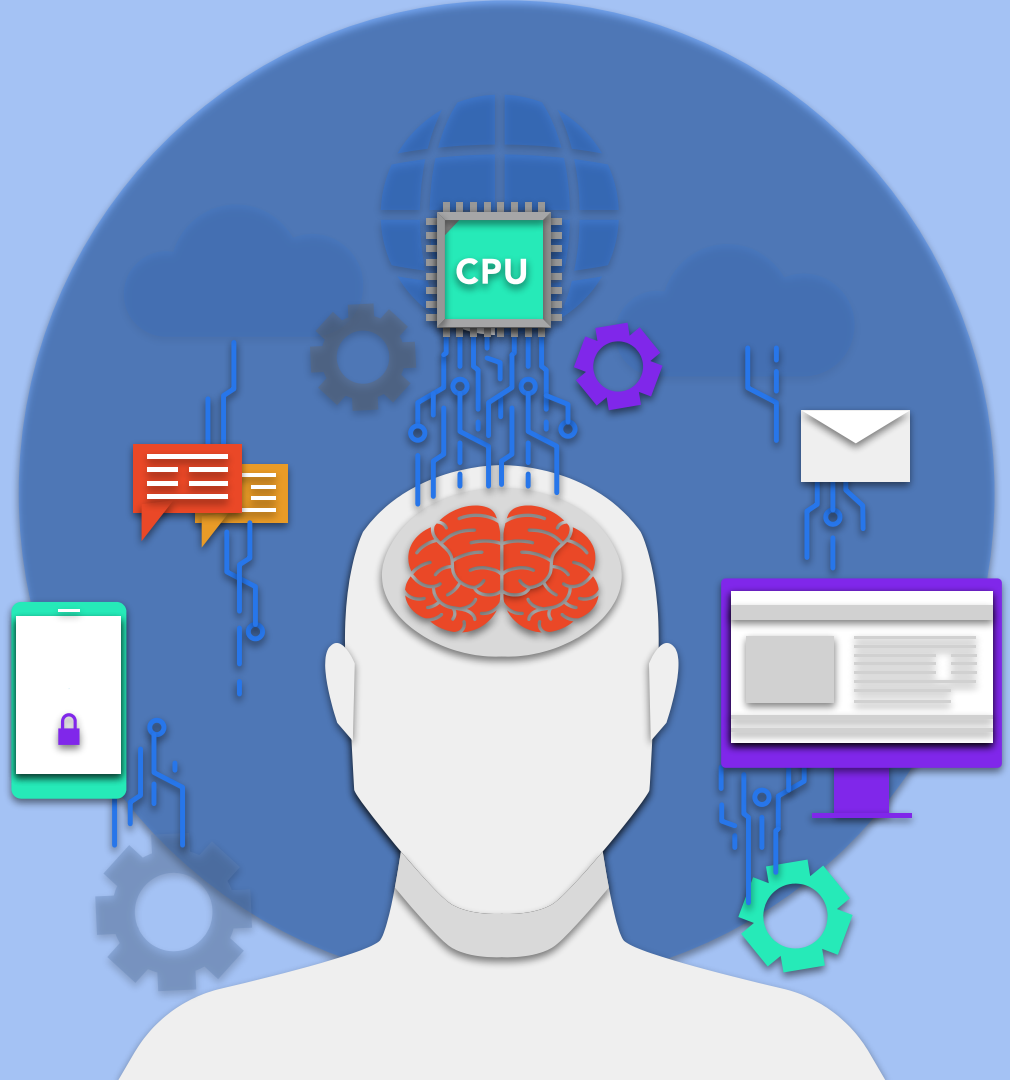


Child Mind Institute - Problematic Internet Use

Members
ABHINAV TRIPATHI
ASHWIN SRIVASTAVA
HARESH AR
MAHALAKSHMI P
MAHALAKSHMI S G
PIYUSHA DHAVALE

Problem Statement

Problematic internet use among children and adolescents is linked to mental health issues but is often difficult to measure due to reliance on professional assessments. Physical activity and fitness data, however, are easily accessible and can indicate early signs of excessive screen time. This project aims to develop a predictive model that uses children's physical activity data to identify early indicators of problematic internet use, enabling timely interventions to promote healthier digital habits.

During participation in the HBN study, some participants were given an accelerometer to wear for up to 30 days continually while at home and going about their regular daily lives.

Severity Impairment Index (sii) [TARGET]

The target sii for this competition is derived from Parent-Child Internet Addiction Test

Sii values: 0 for None, 1 for Mild, 2 for Moderate, and 3 for Severe.

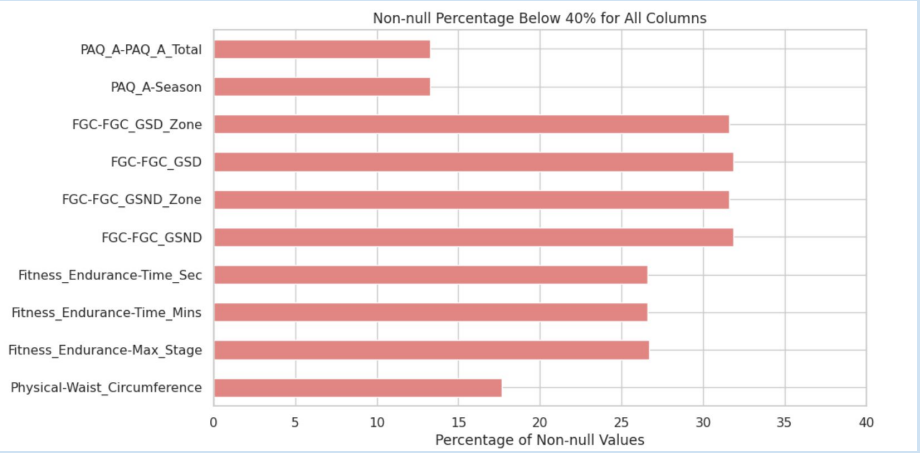
Parent-Child Internet Addiction Test - 20-item scale that measures characteristics and behaviors associated with compulsive use of the Internet including compulsivity, escapism, and dependency. And Total Score is from 0 to 100

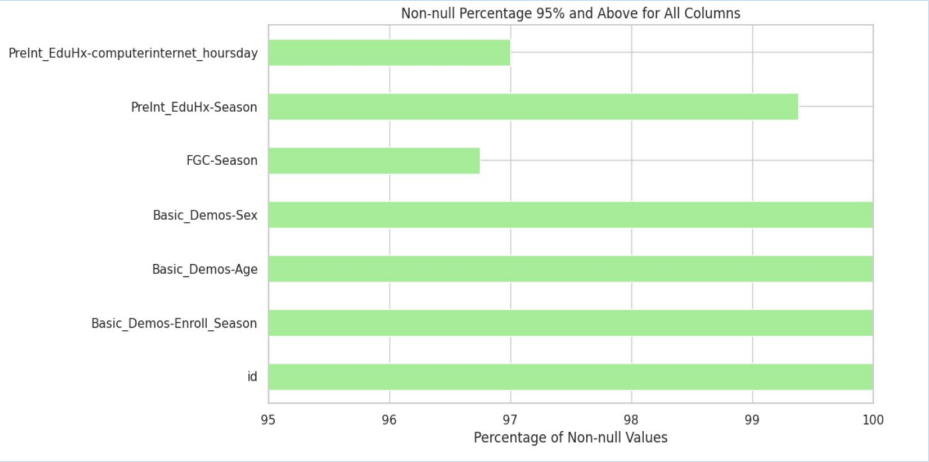
SII	PCIAT
0	0-30
1	31-49
2	50-79
3	80-100

```
'How often does your child disobey time limits you set for online use?',
'How often does your child neglect household chores to spend more time online?',
'How often does your child prefer to spend time online rather than with the rest of your family?',
'How often does your child form new relationships with fellow online users?',
'How often do you complain about the amount of time your child spends online?',
"How often do your child's grades suffer because of the amount of time he or she spends online?",
'How often does your child check his or her e-mail before doing something else?',
'How often does your child seem withdrawn from others since discovering the Internet?',
'How often does your child become defensive or secretive when asked what he or she does online?',
'How often have you caught your child sneaking online against your wishes?',
'How often does your child spend time along in his or her room playing on the computer?',
'How often does your child receive strange phone calls from new "online" friends?',
'How often does your child snap, yell, or act annoyed if bothered while online?',
'How often does your child seem more tired and fatigued than he or she did before the Internet came along?',
'How often does your child seem preoccupied with being back online when off-line?',
'How often does your child throw tantrums with your interference about how long he or she spends online?',
'How often does your child choose to spend time online rather than doing once enjoyed hobbies and/or outside interests?',
'How often does your child become angry or belligerent when your place time limits on how much time he or shes is allowed to spend online?',
'How often does your child choose to spend more time online than going out with friends?',
'How often does your child feel depressed, moody, or nervous when off-line which seems to go away once back online?',
```

Exploratory Data Analysis (EDA)

- **Dataset inspection: Reviewed dimensions, column names, data types, and distributions.**
- **Train-test alignment: Identified and resolved mismatched columns between training and testing datasets.**
- **Missing value handling: Dropped columns with over 40% missing values and retained those with at least 95% non-null data, reducing the dataset to 36 features.**
- **Feature selection: Analyzed correlations and removed redundant columns.**
- **Data transformation: Scaled and normalized features for consistency.**

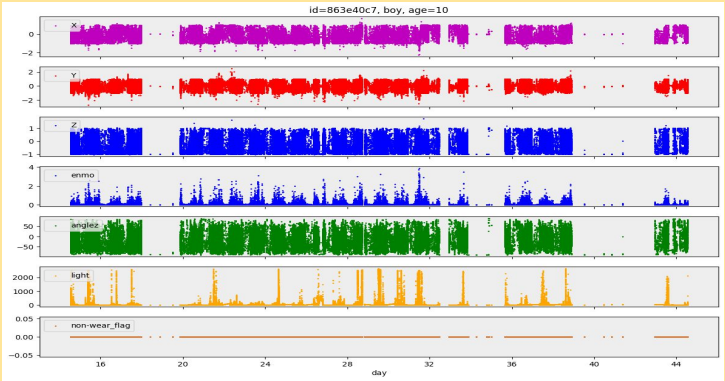


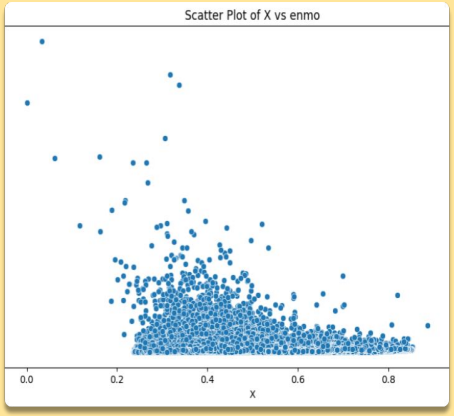


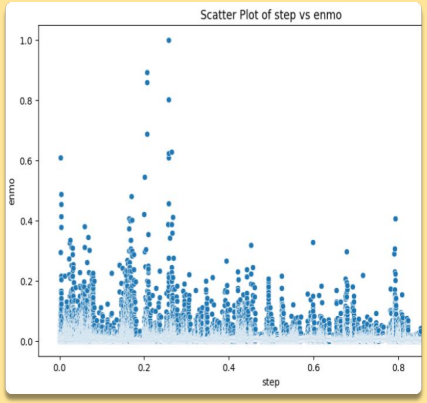
Physical Activity [Actigraphy] Data Analysis

Euclidean Norm Minus One (ENMO) is a metric used to calculate physical activity from accelerometer data and is a important feature of the Actigraphy Dataset. It indicates:

- Physical activity volume (average acceleration)
- Intensity distribution (intensity gradient)
- Intensity of the most active periods (MX metrics) of the day
- Zero values are indicative of periods of no motion.



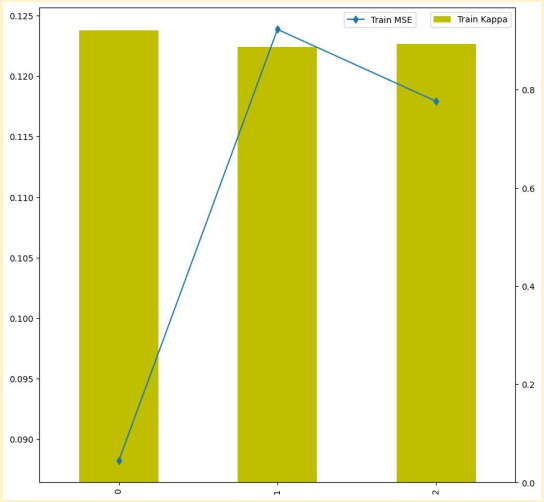


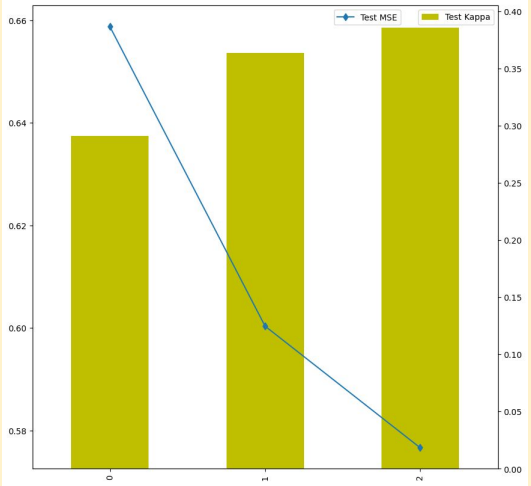


XGB Regression Models

Objective:
To predict PCIAT Score (0 to 100) and then convert to SII

Model #	Features	Hyperparameters	Scaling & Split
Model 1	All Features without EDA	Default	No Scaling, 80:20
Model 2	EDA Features without Categorical Season	learning_rate, max_depth, Min_child_weight,n_estimators Least Standard deviation in test score among splits.	StandardScaler, 80:20
Model 3	EDA Features		



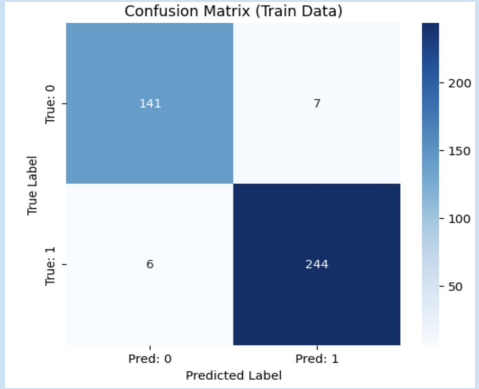


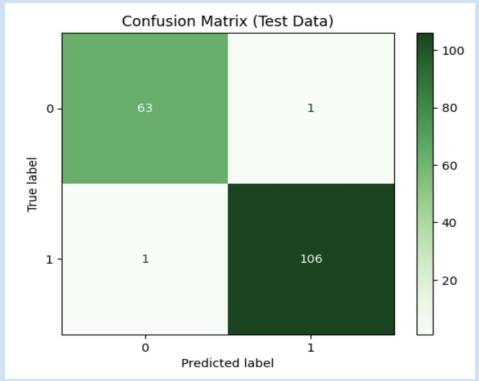
Logistic Regression Model

- This model was developed using exploratory data analysis (EDA) features, with data scaling and splitting also performed

[231]:
```
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
```

```
Accuracy: 0.9883
Precision: 0.9907
Recall: 0.9907
```





Things tried but not completed

- PCA
- LightGBM Regressor
- Random Forest Classification
- K-Means Clustering

PCA (PRINCIPAL COMPONENT ANALYSIS)

Data Overview:

- Dataset loaded using Pandas.
- Took Target variable: **‘sii'**

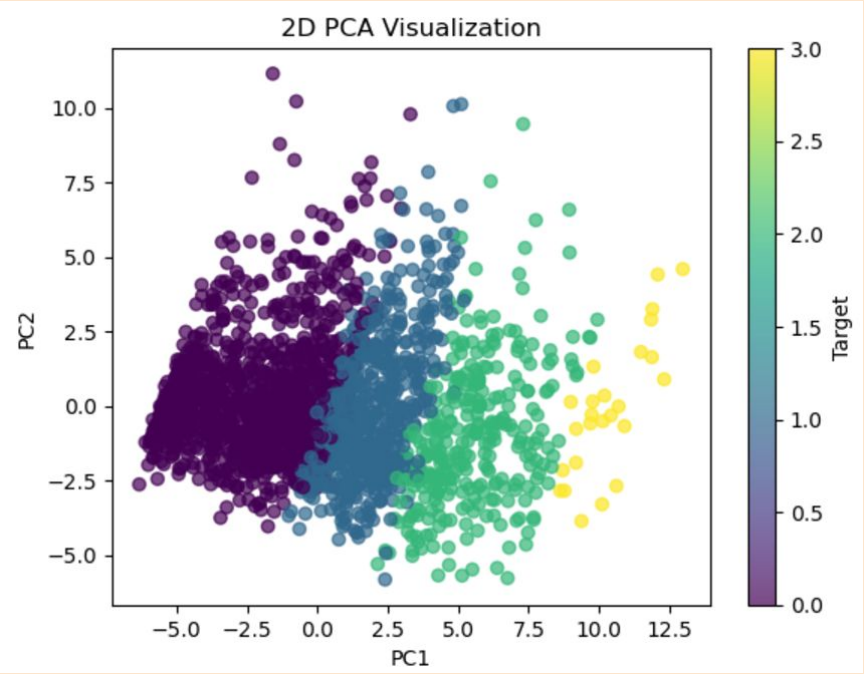
Data Pre-processing Steps:

- Handled missing values by dropping rows with missing target values
- Separated numerical and categorical features
- Applied preprocessing pipelines:

1. Numerical Data: Imputed with mean, then scaled
2. Categorical Data: Imputed with mode, then one-hot encoded

Dimensionality Reduction in PCA

- Explored variance explained by each principal component
- Selected components that cumulatively explain significant variance
- Visualized the reduction process



1. Variance explained by **PC1**: **14.17%**
2. Variance explained by **PC2**: **11.42%**
3. Total variance explained by **PC1** and **PC2**: **25.59%**
4. Target variable **(sii)** ranges from **0.0** to **3.0**
5. Distinct clusters suggests that there may be groups of data with similar sii values. - The first two components capture a significant portion of the variance, but further analysis on more components may be needed for deeper insights.

Contributions

Task	Members
Data Cleanup	ABHINAV TRIPATHI ASHWIN SRIVASTAVA MAHALAKSHMI P PIYUSHA DHAVALE
ML Model Selection/ Training	ASHWIN SRIVASTAVA HARESH AR MAHALAKSHMI P MAHALAKSHMI S G PIYUSHA DHAVALE
Hyper parameter tuning and Metrics	HARESH AR PIYUSHA DHAVALE
Presentation	ALL MEMBERS
Documentation	ALL MEMBERS

Conclusion

- XGBoost Model Performance improved as we applied standard scaler, feature selection and hyperparameter tuning.
- The logistic regression model showed improved performance in terms of accuracy, precision, and recall as reflected in the updated confusion matrix.
- Next Scope:
  - Voting module for different models in the same notebook that would help in choosing the right model for prediction.