Problematic Internet Use Classification

Michael Lu
October 25, 2024
Brown University, Department of Computer Science
https://github.com/mdlu02/InternetUseClassification

Introduction

Problem

- Increased internet use associated with mental health problems.
- Especially prevalent among young individuals.

Data

- Healthy Brain Network (HBN) dataset
- ~5000 5-22 year-olds selected after clinical and research screenings.
- Tabular and time series data

Goal

Build a classification model to predict the `sii` (Severity Impairment Index) of a given individual.

- `sii` is based on buckets of `PCIAT-PCIAT_Total` scores

Classification vs. Regression?

Challenges

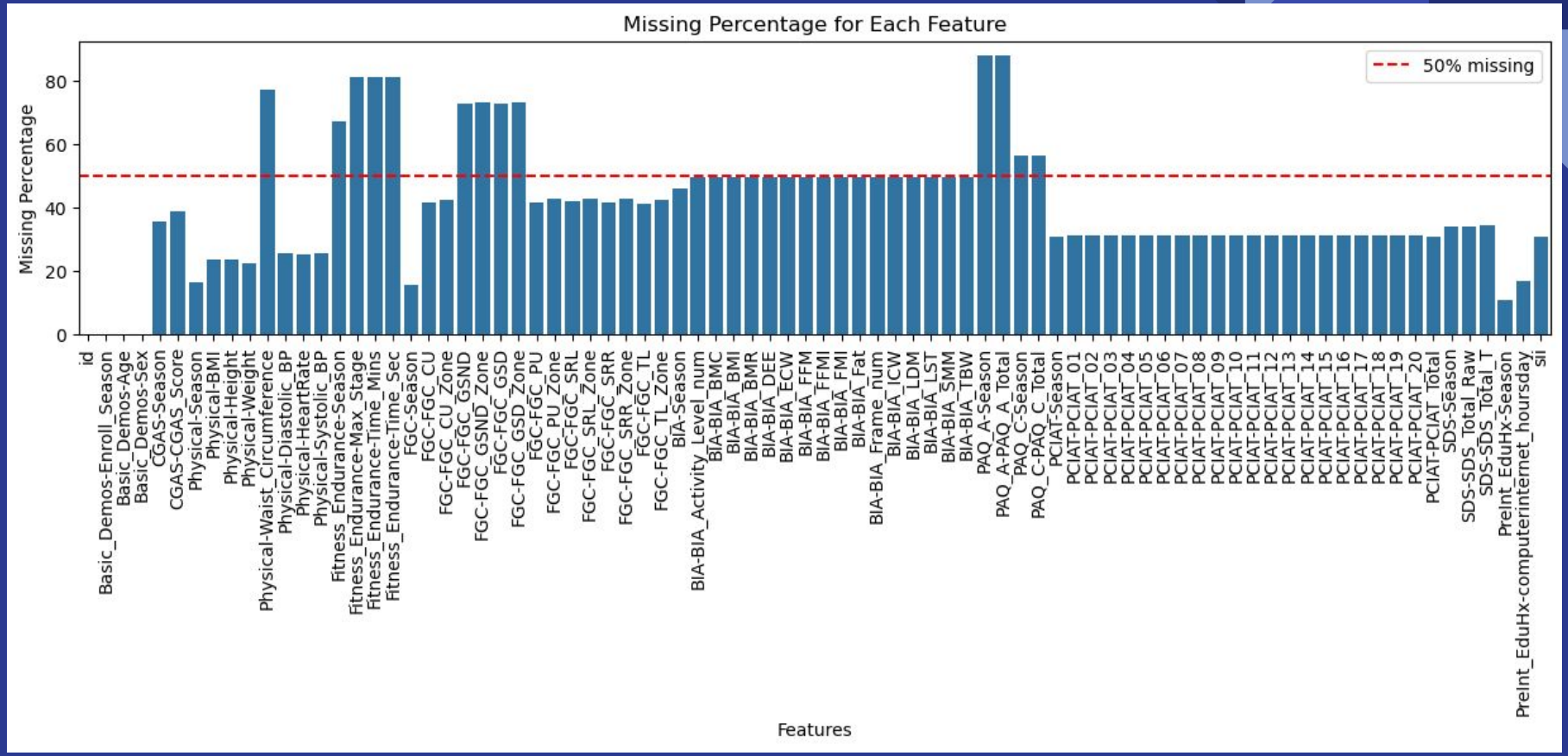
Missing Data Time Series Data

Missing tabular data and 74.82% do not have activity tracker time series data.

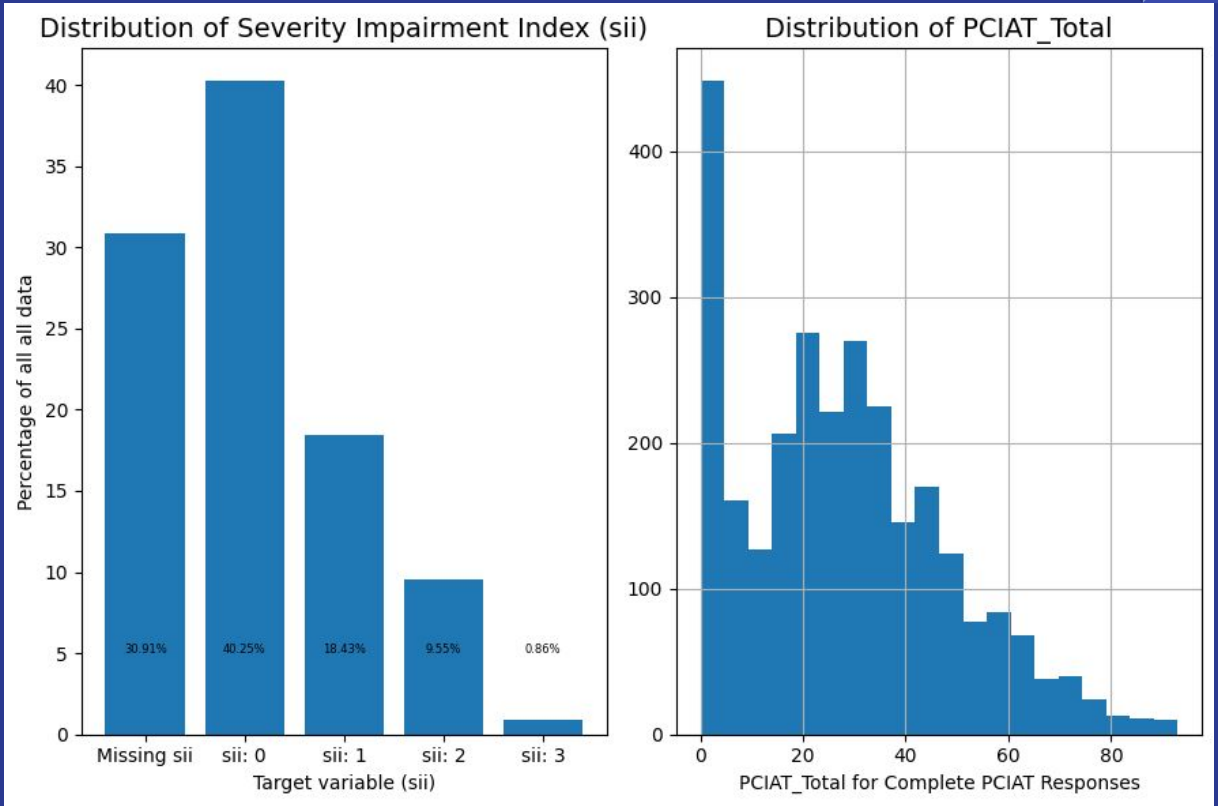
Missing Label Data

30.9% of children are missing `PCIAT-PCIAT_Total` and `sii` data.

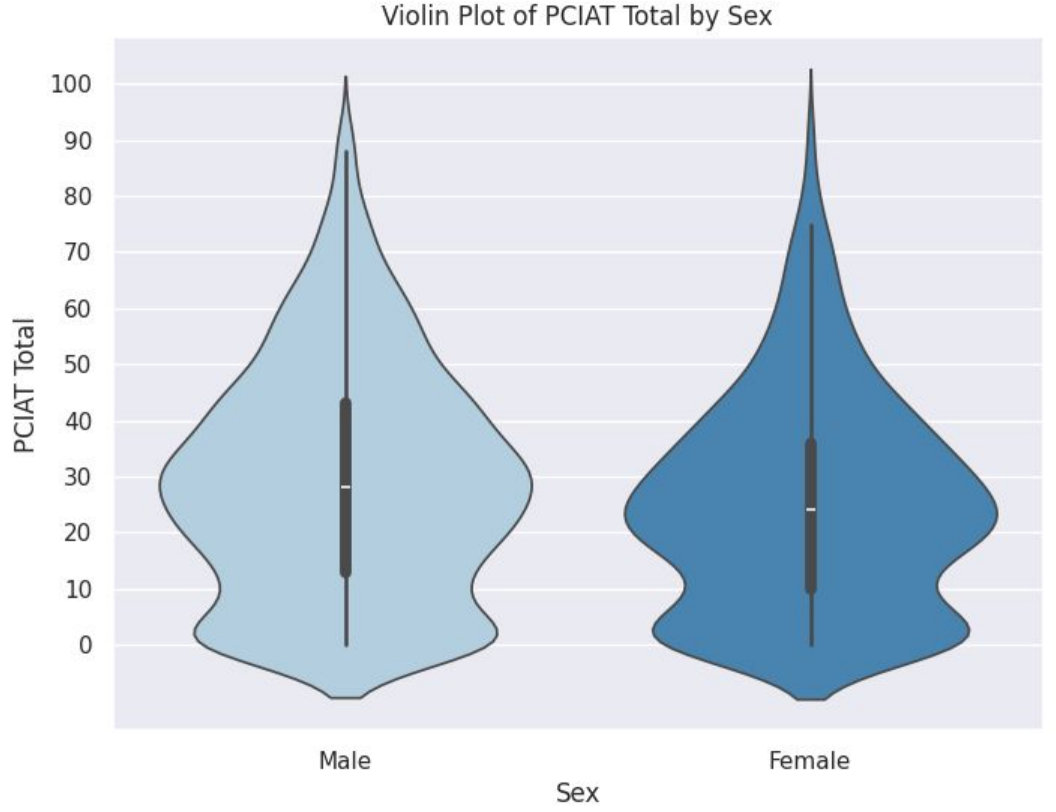
Missing Data



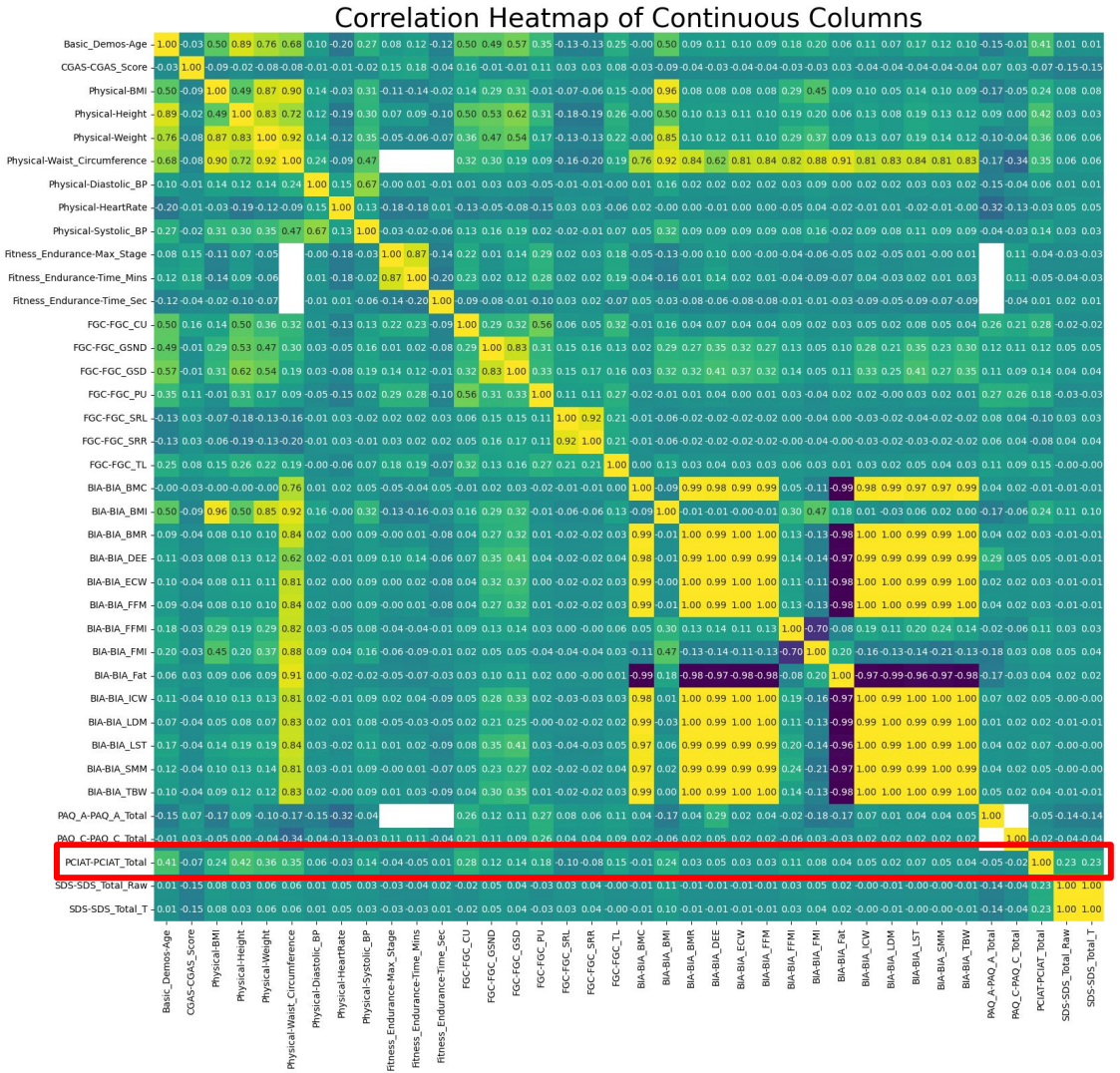
Target variable(s)



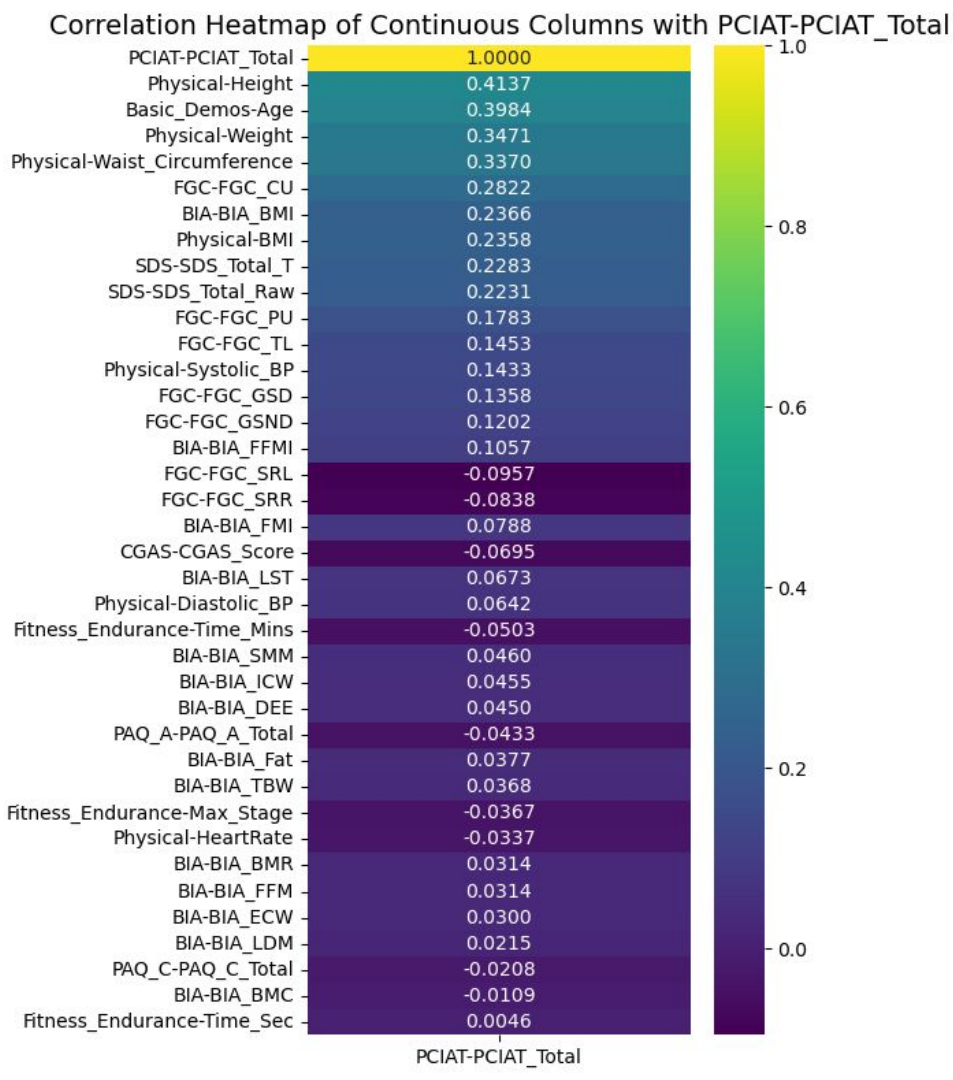
Target variable(s)



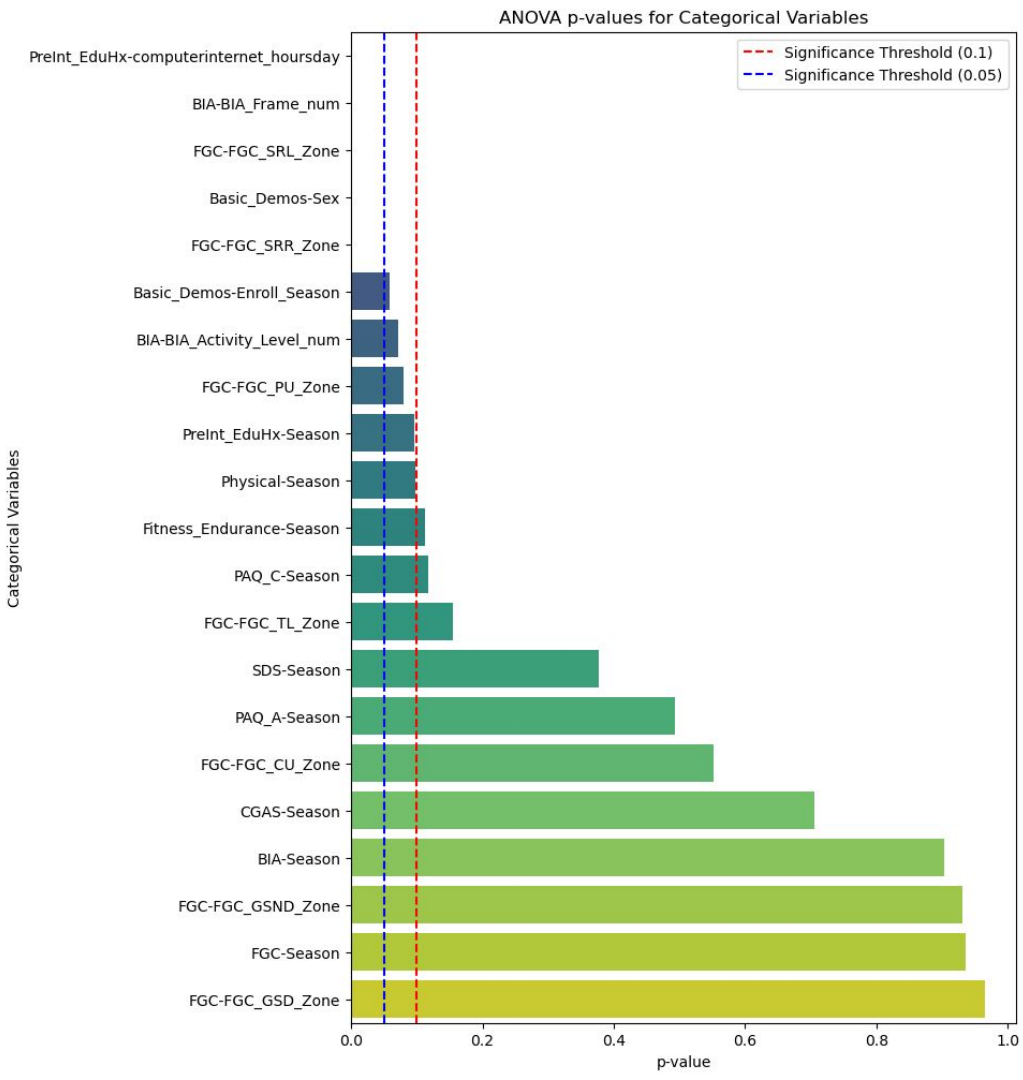
Correlation of continuous features vs PCIAT-PCIAT_Total



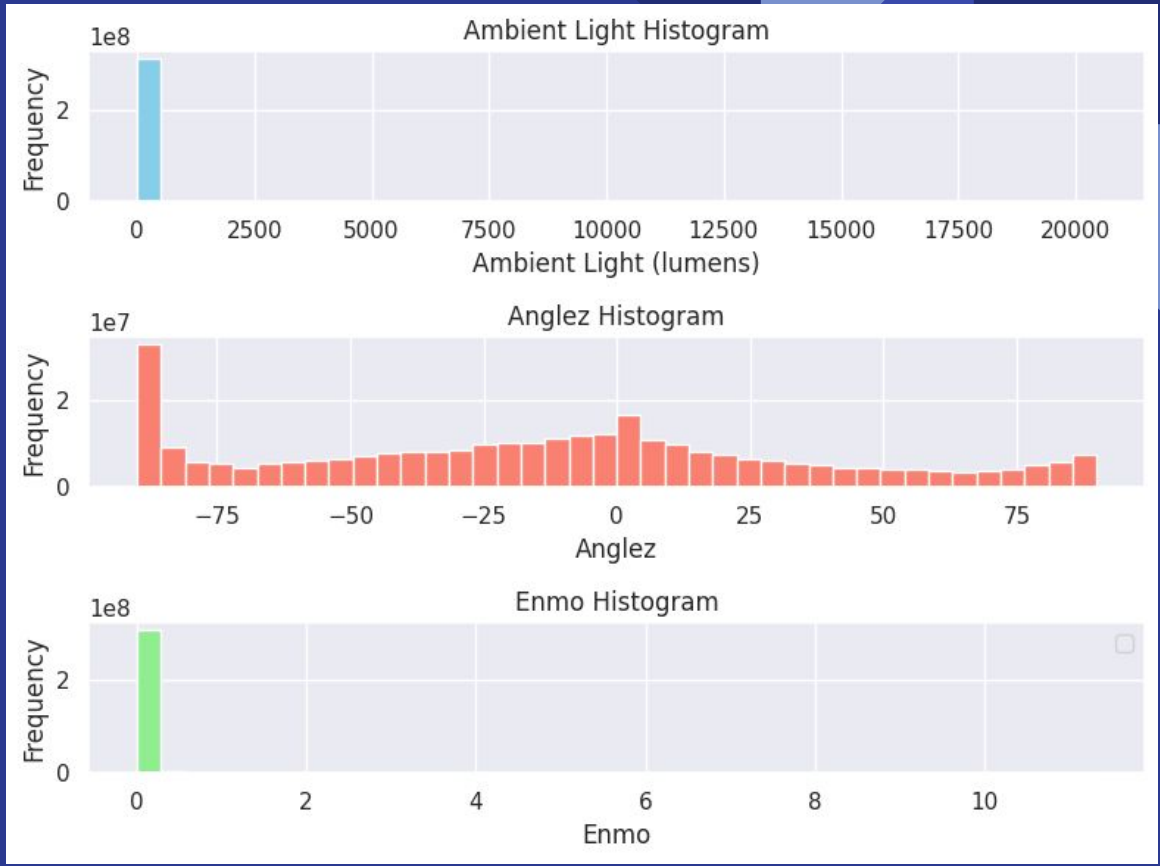
Correlation of continuous features vs PCIAT-PCIAT_Total



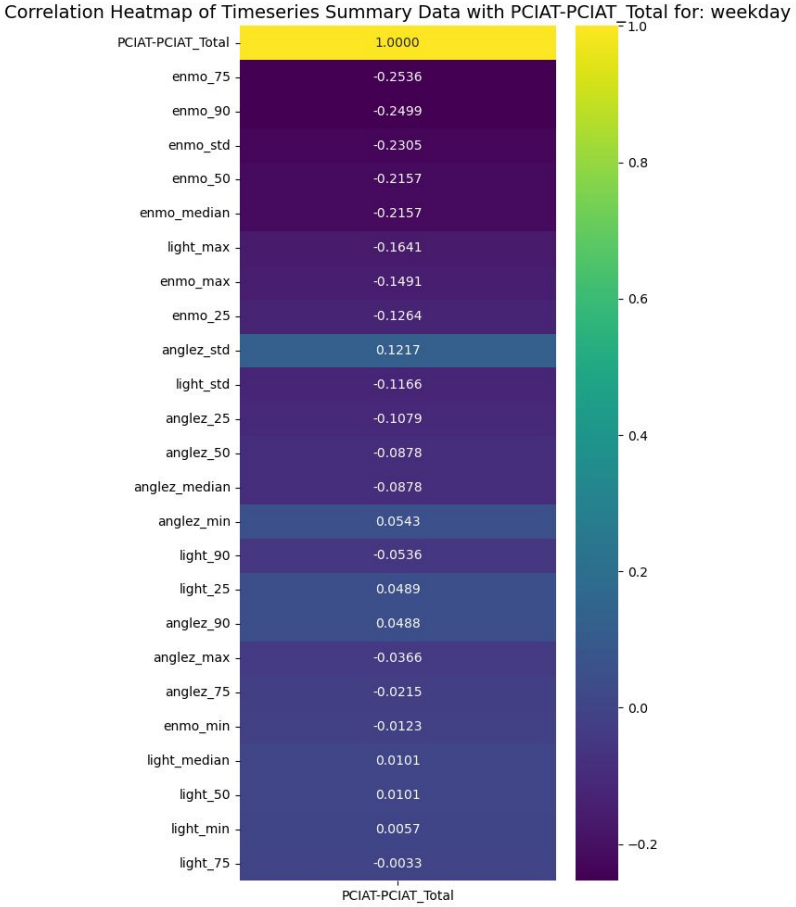
Welch ANOVA results on categorical features vs. PCIAT-PCIAT_Total (p-values)

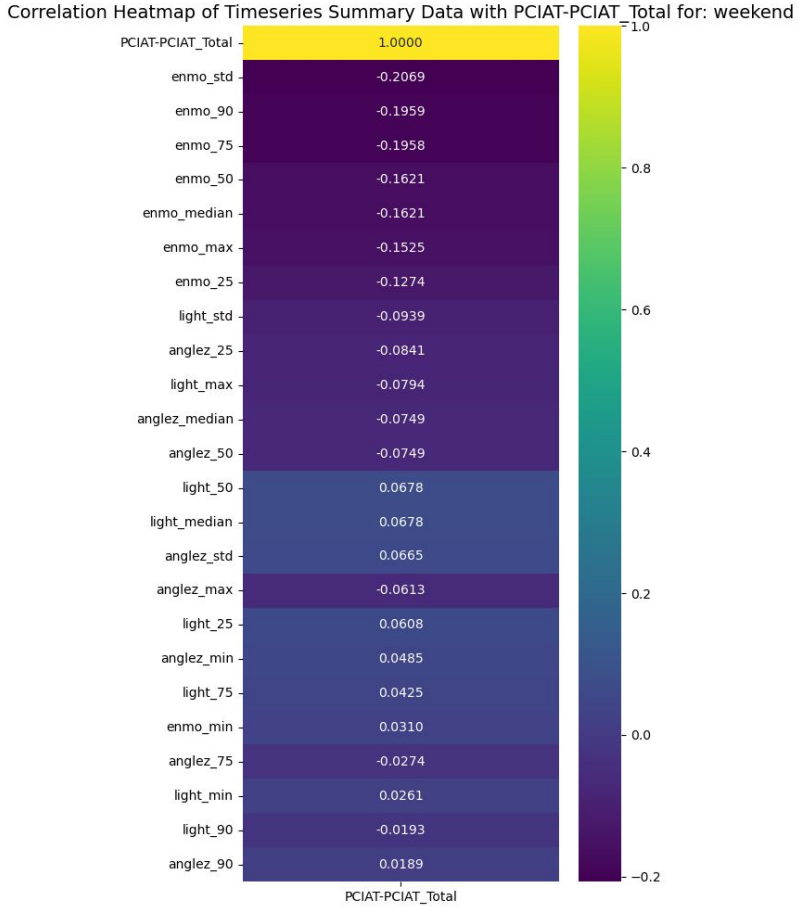


Time Series Summary Metrics vs. PCIAT-PCIAT_Total



Time series data (anglez, light, and enmo) summary statistics correlations vs. PCIAT-PCIAT_Total for weekdays and weekends





Split

Using: `train_test_split` and `StratifiedKFold`

- Grouping?
  - Not necessary for task
- Stratification?
  - Existence of time series data, sex, age, and `sii` classification
- Standard `KFold` splitting

Preprocess

Features

- Drop data with missing target variables
- Remove highly correlated feature pairs
- Remove insignificant categorical features
- Impute missing data based on feature type

Preprocess

Transformations

- `OneHotEncoder` for categorical features
- `OrdinalEncoder` for ordinal features
- Use `MinMaxScaler` and `StandardScaler` for continuous features depending on existence of upper and lower bounds.

Preprocess

Overview

- Features before preprocessing: **95** (82 tabular and 13 time series)
- Features after preprocessing: **43±1** based on feature selection for the fold
- Train, validation, test split sizes after 3-fold split: **~519, ~260, 199** respectively