EC 9560 – DATA MINING

LAB 01

DARMILA.T

2020/E/027

SEMESTER 07

02th OCTOBER 2024

Title: Predict the level of Problematic Internet Use Among Children and Adolescents Using based on demographic, physical health, internet use, and fitness data.

Objective: To predict the level of problematic internet usage exhibited by children and adolescents based on their physical activity and fitness data. By developing a predictive model, the aim is to identify early indicators of problematic internet use, allowing for timely interventions that promote healthier digital habits.

Methodology:

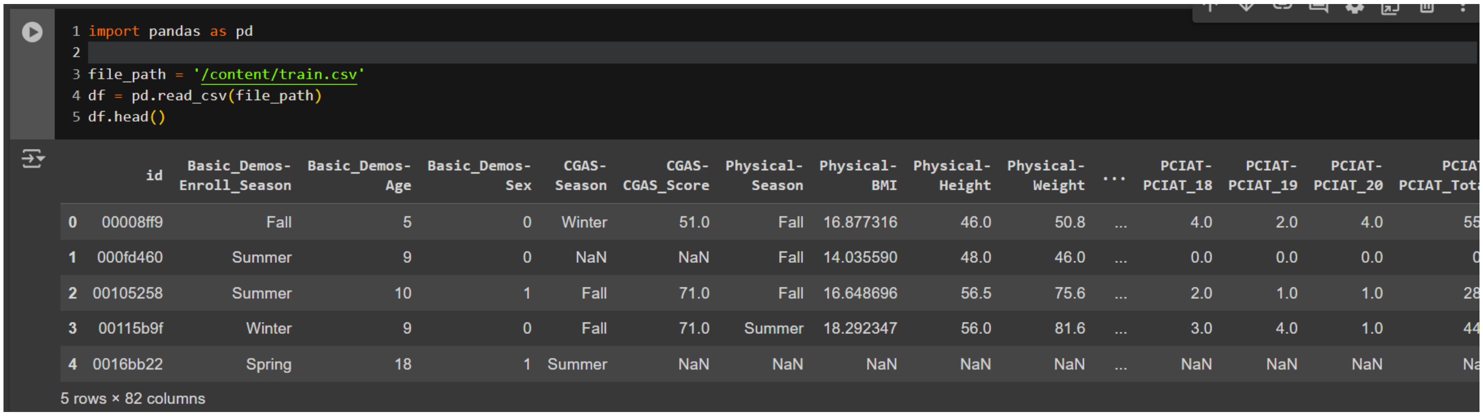
1. Data Collection
   - Load and Explore the Dataset: Start by loading the train.csv file, which contains features such as demographic information, internet usage behavior, fitness measures, and other health indicators. Analyze the data_dictionary.csv to fully understand the meaning and context of each field.
2. Data Preprocessing
   - Handle Missing Data: Address missing values using techniques like mean or median imputation, or consider removing features that have an excessive amount of missing data.
   - Feature Engineering: Create new features if they can provide additional insight or improve model performance.
   - Data Cleaning: Identify and resolve invalid or inconsistent data points. Convert categorical variables into numerical representations using methods like one-hot encoding or label encoding.
3. Data Splitting
   - Train-Validation Split: Divide the preprocessed training data into training and validation sets to enable model tuning and evaluation.
   - Stratified Sampling: Since the target variable (Severity Impairment Index, sii) is ordinal with categories (None, Mild, Moderate, Severe), use stratified sampling to preserve the class distribution during the split.
4. Model Selection
   - Algorithms for Classification:
     - Logistic Regression
     - Decision Trees
     - k-Nearest Neighbors (KNN)
     - Naive Bayes
     - Support Vector Machines (SVM)

    - Random Forest Classifier
    - Gradient Boosting Classifier

5. Model Training and Hyperparameter Tuning
    - Model Training: Train several models on the training set and use cross-validation to prevent overfitting.
    - Hyperparameter Optimization: Utilize hyperparameter tuning techniques like Grid Search or Random Search to find the optimal configuration for each model.
6. Evaluation
    - Metrics for Evaluation:
        - Accuracy: Measure how often the model predicts the correct category.
        - Precision, Recall, and F1-Score: These metrics are important for assessing performance, especially with imbalanced classes.
        - Confusion Matrix: Use this to visualize misclassifications among the different severity levels.
        - AUC-ROC Curve: Evaluate the model's ability to distinguish between the different severity levels of problematic internet use.
7. Final Model and Predictions
    - Once the best-performing model has been finalized, apply it to the test.csv dataset (test set) to predict the Severity Impairment Index (sii).

Data description with a link to data in data repository:

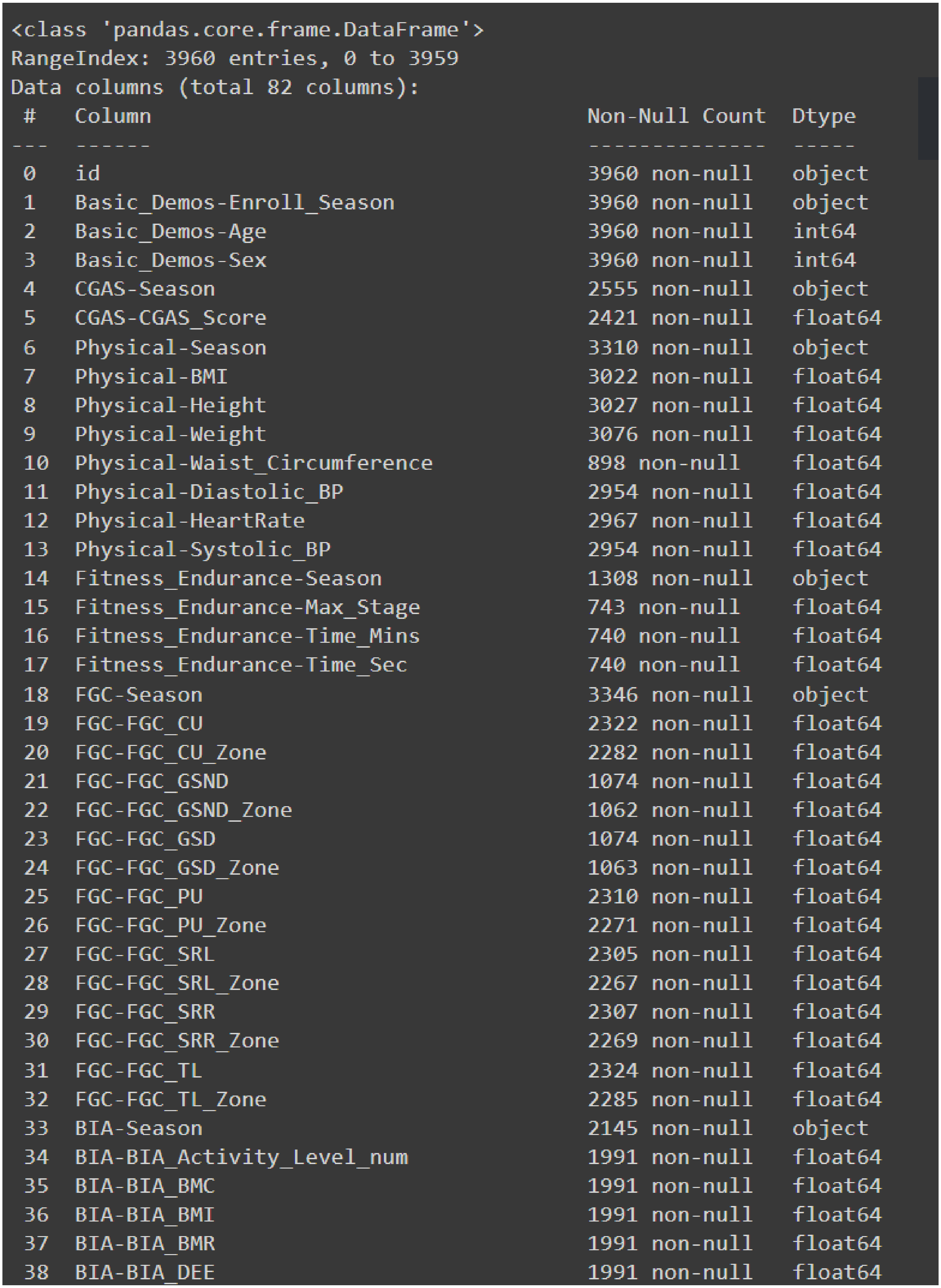
https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use/data?select=train.csv



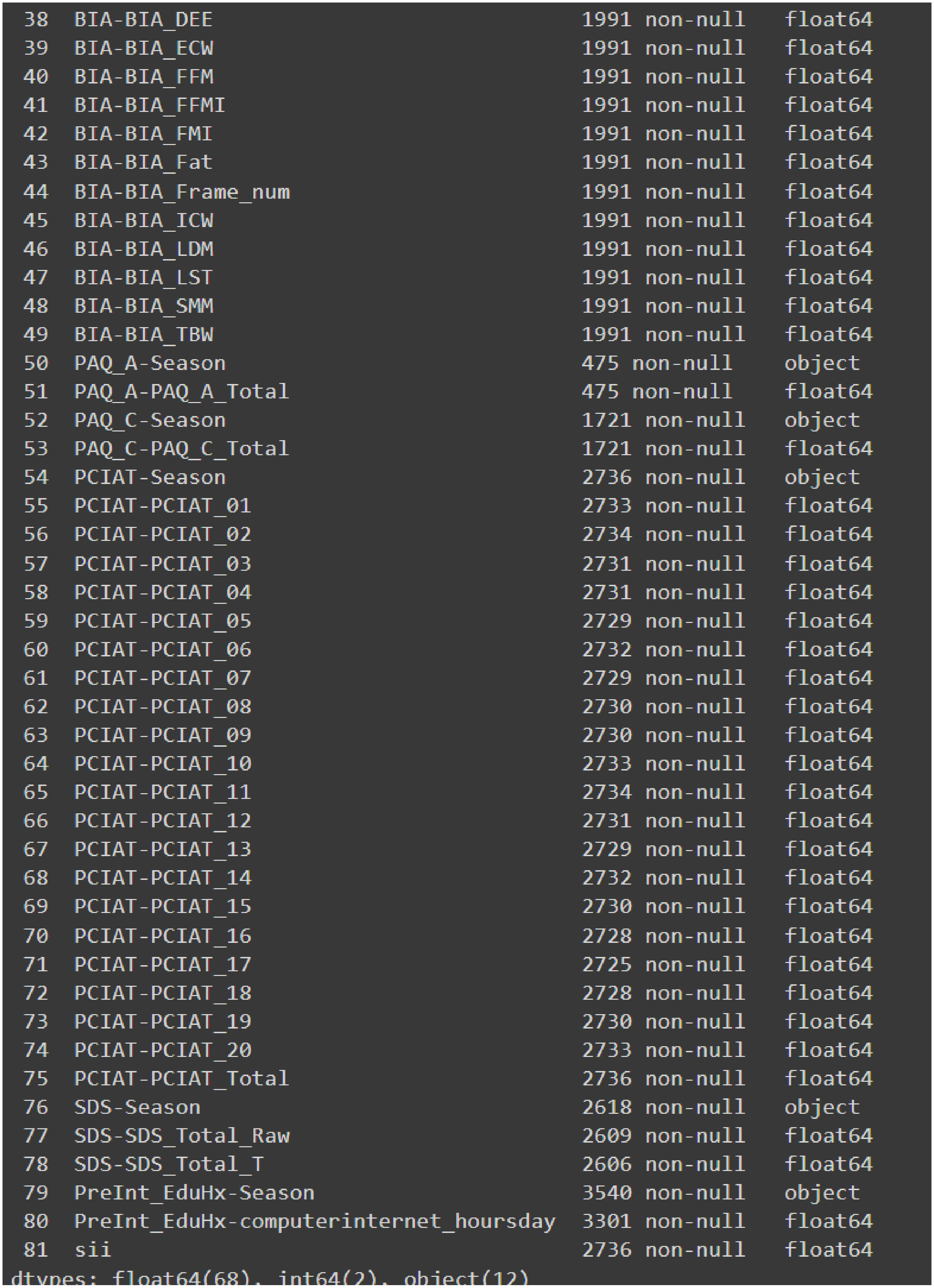
```python
import pandas as pd

file_path = '/content/train.csv'
df = pd.read_csv(file_path)
df.head()
```

	id	Basic_Demos-Enroll_Season	Basic_Demos-Age	Basic_Demos-Sex	CGAS-Season	CGAS-CGAS_Score	Physical-Season	Physical-BMI	Physical-Height	Physical-Weight	...	PCIAT-PCIAT_18	PCIAT-PCIAT_19	PCIAT-PCIAT_20	PCIA PCIAT_Tot
0	00008ff9	Fall	5	0	Winter	51.0	Fall	16.877316	46.0	50.8	...	4.0	2.0	4.0	55
1	000fd460	Summer	9	0	NaN	NaN	Fall	14.035590	48.0	46.0	...	0.0	0.0	0.0	0
2	00105258	Summer	10	1	Fall	71.0	Fall	16.648696	56.5	75.6	...	2.0	1.0	1.0	28
3	00115b9f	Winter	9	0	Fall	71.0	Summer	18.292347	56.0	81.6	...	3.0	4.0	1.0	44
4	0016bb22	Spring	18	1	Summer	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	Na

5 rows × 82 columns



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3960 entries, 0 to 3959
Data columns (total 82 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   id                                      3960 non-null   object
 1   Basic_Demos-Enroll_Season               3960 non-null   object
 2   Basic_Demos-Age                         3960 non-null   int64
 3   Basic_Demos-Sex                         3960 non-null   int64
 4   CGAS-Season                             2555 non-null   object
 5   CGAS-CGAS_Score                         2421 non-null   float64
 6   Physical-Season                         3310 non-null   object
 7   Physical-BMI                            3022 non-null   float64
 8   Physical-Height                         3027 non-null   float64
 9   Physical-Weight                         3076 non-null   float64
 10  Physical-Waist_Circumference            898 non-null    float64
 11  Physical-Diastolic_BP                   2954 non-null   float64
 12  Physical-HeartRate                      2967 non-null   float64
 13  Physical-Systolic_BP                    2954 non-null   float64
 14  Fitness_Endurance-Season                1308 non-null   object
 15  Fitness_Endurance-Max_Stage             743 non-null    float64
 16  Fitness_Endurance-Time_Mins             740 non-null    float64
 17  Fitness_Endurance-Time_Sec              740 non-null    float64
 18  FGC-Season                              3346 non-null   object
 19  FGC-FGC_CU                              2322 non-null   float64
 20  FGC-FGC_CU_Zone                         2282 non-null   float64
 21  FGC-FGC_GSND                            1074 non-null   float64
 22  FGC-FGC_GSND_Zone                       1062 non-null   float64
 23  FGC-FGC_GSD                             1074 non-null   float64
 24  FGC-FGC_GSD_Zone                        1063 non-null   float64
 25  FGC-FGC_PU                              2310 non-null   float64
 26  FGC-FGC_PU_Zone                         2271 non-null   float64
 27  FGC-FGC_SRL                             2305 non-null   float64
 28  FGC-FGC_SRL_Zone                        2267 non-null   float64
 29  FGC-FGC_SRR                             2307 non-null   float64
 30  FGC-FGC_SRR_Zone                        2269 non-null   float64
 31  FGC-FGC_TL                              2324 non-null   float64
 32  FGC-FGC_TL_Zone                         2285 non-null   float64
 33  BIA-Season                              2145 non-null   object
 34  BIA-BIA_Activity_Level_num              1991 non-null   float64
 35  BIA-BIA_BMC                             1991 non-null   float64
 36  BIA-BIA_BMI                             1991 non-null   float64
 37  BIA-BIA_BMR                             1991 non-null   float64
 38  BIA-BIA_DEE                             1991 non-null   float64
```



```
 38  BIA-BIA_DEE                             1991 non-null   float64
 39  BIA-BIA_ECW                             1991 non-null   float64
 40  BIA-BIA_FFM                             1991 non-null   float64
 41  BIA-BIA_FFMI                            1991 non-null   float64
 42  BIA-BIA_FMI                             1991 non-null   float64
 43  BIA-BIA_Fat                             1991 non-null   float64
 44  BIA-BIA_Frame_num                       1991 non-null   float64
 45  BIA-BIA_ICW                             1991 non-null   float64
 46  BIA-BIA_LDM                             1991 non-null   float64
 47  BIA-BIA_LST                             1991 non-null   float64
 48  BIA-BIA_SMM                             1991 non-null   float64
 49  BIA-BIA_TBW                             1991 non-null   float64
 50  PAQ_A-Season                            475 non-null    object 
 51  PAQ_A-PAQ_A_Total                       475 non-null    float64
 52  PAQ_C-Season                            1721 non-null   object 
 53  PAQ_C-PAQ_C_Total                       1721 non-null   float64
 54  PCIAT-Season                            2736 non-null   object 
 55  PCIAT-PCIAT_01                          2733 non-null   float64
 56  PCIAT-PCIAT_02                          2734 non-null   float64
 57  PCIAT-PCIAT_03                          2731 non-null   float64
 58  PCIAT-PCIAT_04                          2731 non-null   float64
 59  PCIAT-PCIAT_05                          2729 non-null   float64
 60  PCIAT-PCIAT_06                          2732 non-null   float64
 61  PCIAT-PCIAT_07                          2729 non-null   float64
 62  PCIAT-PCIAT_08                          2730 non-null   float64
 63  PCIAT-PCIAT_09                          2730 non-null   float64
 64  PCIAT-PCIAT_10                          2733 non-null   float64
 65  PCIAT-PCIAT_11                          2734 non-null   float64
 66  PCIAT-PCIAT_12                          2731 non-null   float64
 67  PCIAT-PCIAT_13                          2729 non-null   float64
 68  PCIAT-PCIAT_14                          2732 non-null   float64
 69  PCIAT-PCIAT_15                          2730 non-null   float64
 70  PCIAT-PCIAT_16                          2728 non-null   float64
 71  PCIAT-PCIAT_17                          2725 non-null   float64
 72  PCIAT-PCIAT_18                          2728 non-null   float64
 73  PCIAT-PCIAT_19                          2730 non-null   float64
 74  PCIAT-PCIAT_20                          2733 non-null   float64
 75  PCIAT-PCIAT_Total                       2736 non-null   float64
 76  SDS-Season                              2618 non-null   object 
 77  SDS-SDS_Total_Raw                       2609 non-null   float64
 78  SDS-SDS_Total_T                         2606 non-null   float64
 79  PreInt_EduHx-Season                     3540 non-null   object 
 80  PreInt_EduHx-computerinternet_hoursday  3301 non-null   float64
 81  sii                                     2736 non-null   float64
dtypes: float64(68), int64(2), object(12)
```

Reference :

https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use/overview