Child Mind Institute: Problematic Internet Use

Relating Physical Activity to Problematic Internet Use
Evaluation and Optimization of Classification Models

Submitted by: Aishwarya Malhotra
Under the Guidance of: Prof. Eugene Pinsky

Objective

Predict the Severity Impairment Index (sii) to assess Problematic Internet Use (PIU) in children and young people, various classification models.

Data Overview:

- **Train Data:** 3,960 records with 81 features (excluding ID).
- **Target Variable**: si
- **Test Data**: Formatted sample with 58 columns; actual hidden test set includes ~3,800 instances.

Missing Data: Significant missing data challenges exist:

- Over **100,000** values missing across the dataset.
- **1,224** records lack both the target and all PCIAT fields.
- Only **2,736** records include the target variable.

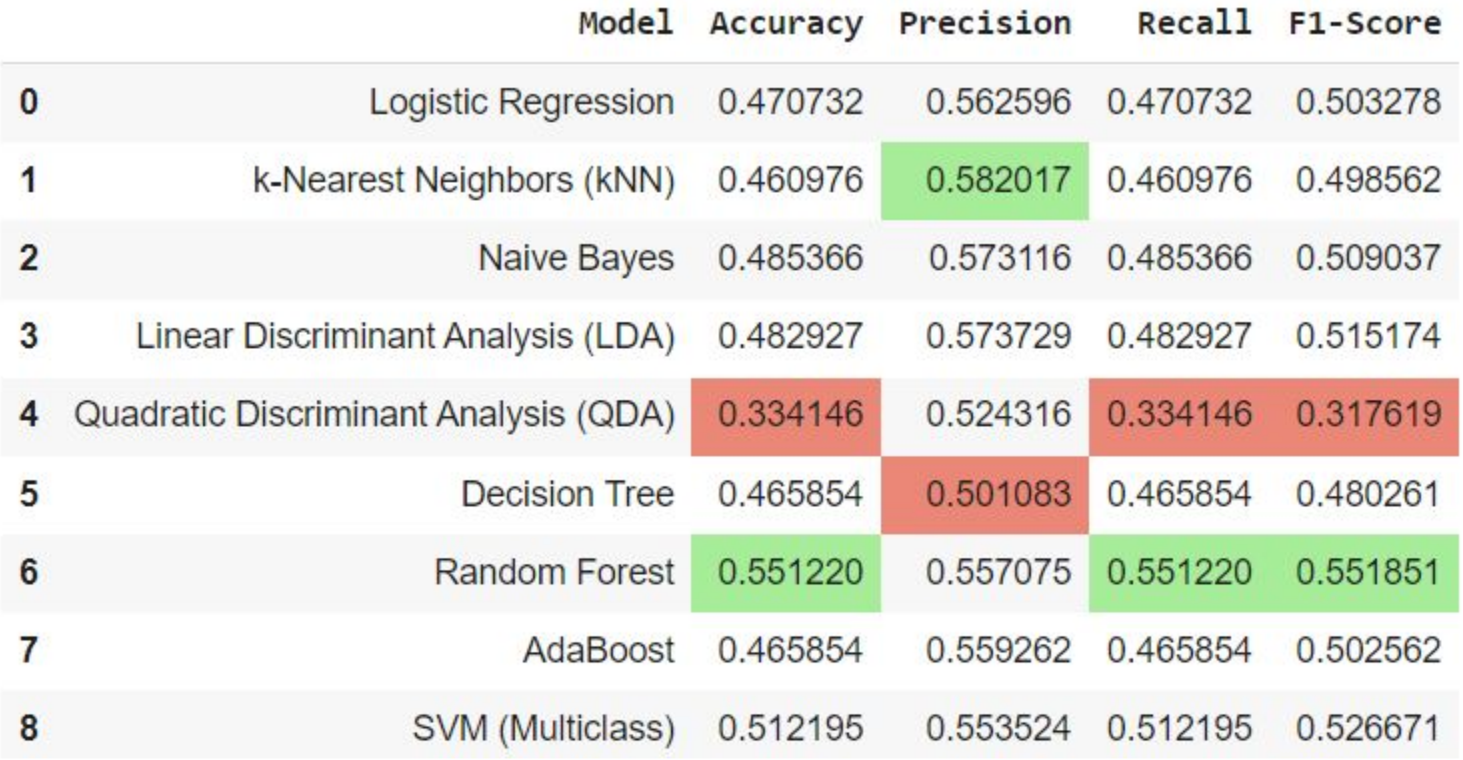
Models Explored:

1. Logistic Regression
2. k-Nearest Neighbors (kNN)
3. Naive Bayes
4. Decision Trees and Random Forest
5. Linear Discriminant Analysis (LDA)
6. Quadratic Discriminant Analysis (QDA)
7. Support Vector Machines (SVM)
8. AdaBoost

Project Prompts - Classification

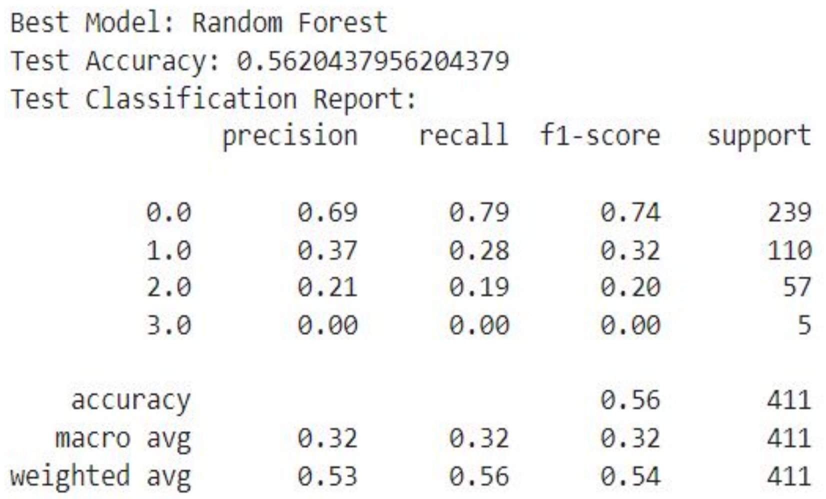
1. **Dropping Columns which are not needed for this project**
2. **EDA**
3. **Correlation**
4. **Handling Missing Values**
5. **Feature and Label Extraction**
6. **Data Preprocessing**
   a. Handling Missing Values
   b. Feature Normalization
7. **Data Splitting (70/30)**
8. **Model Training**
   a. Classification: Logistic regression, k-NN, Naive Bayes, LDA, QDA, Decision tree, Random forest, AdaBoost, and SVM
   b. Regression: Linear Regression, kNN Regressor, Random Forest Regressor, Gradient Boosting Regressor.
9. **Model Evaluation**
10. **Hyperparameter Tuning on the Best Model**
11. **Optimization and Final Evaluation**
12. **Predictions on Test Set**

Model Comparison

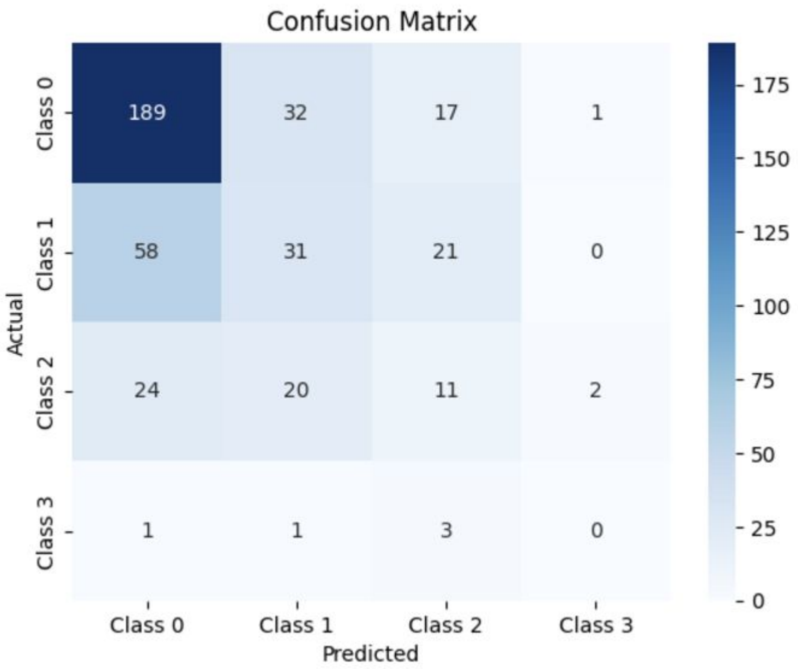


	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.470732	0.562596	0.470732	0.503278
1	k-Nearest Neighbors (kNN)	0.460976	0.582017	0.460976	0.498562
2	Naive Bayes	0.485366	0.573116	0.485366	0.509037
3	Linear Discriminant Analysis (LDA)	0.482927	0.573729	0.482927	0.515174
4	Quadratic Discriminant Analysis (QDA)	0.334146	0.524316	0.334146	0.317619
5	Decision Tree	0.465854	0.501083	0.465854	0.480261
6	Random Forest	0.551220	0.557075	0.551220	0.551851
7	AdaBoost	0.465854	0.559262	0.465854	0.502562
8	SVM (Multiclass)	0.512195	0.553524	0.512195	0.526671

Model Evaluation & Deployment: Random Forest

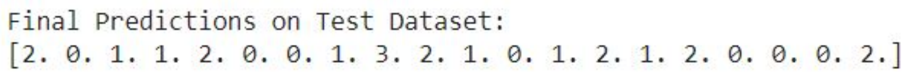


```
Best Model: Random Forest
Test Accuracy: 0.5620437956204379
Test Classification Report:
              precision    recall  f1-score   support

         0.0       0.69      0.79      0.74       239
         1.0       0.37      0.28      0.32       110
         2.0       0.21      0.19      0.20        57
         3.0       0.00      0.00      0.00         5

    accuracy                           0.56       411
   macro avg       0.32      0.32      0.32       411
weighted avg       0.53      0.56      0.54       411
```



	Class 0	Class 1	Class 2	Class 3
Class 0	189	32	17	1
Class 1	58	31	21	0
Class 2	24	20	11	2
Class 3	1	1	3	0

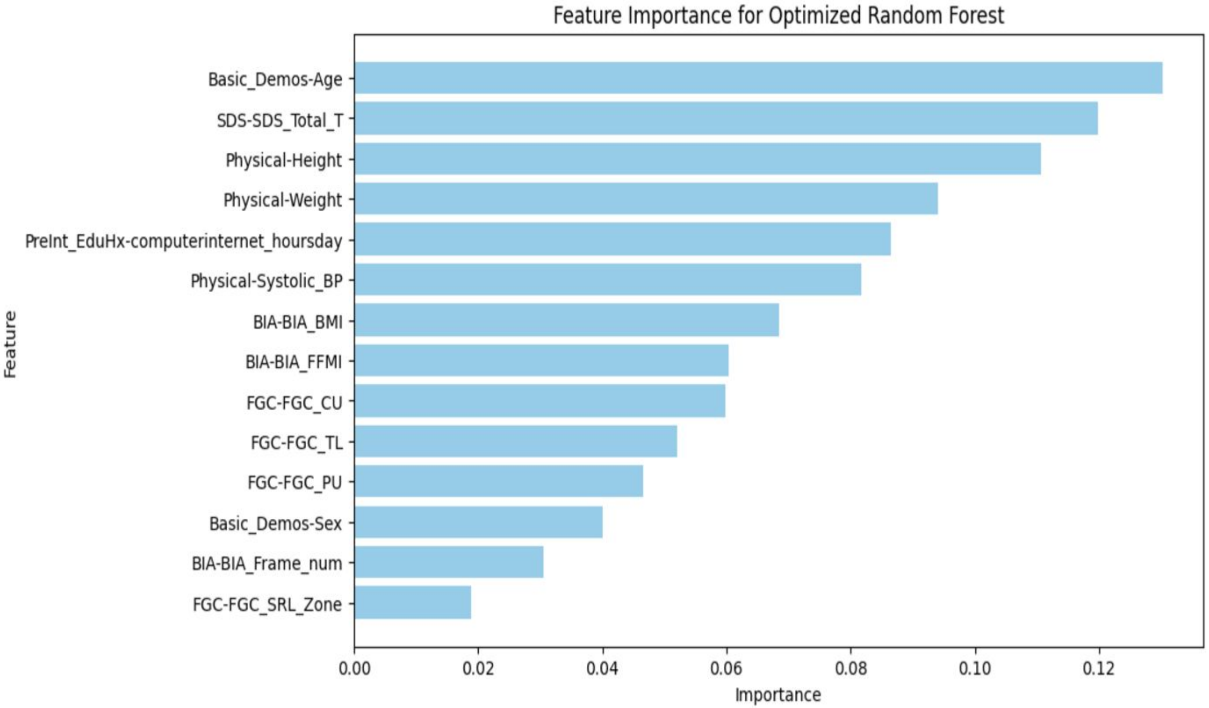
Model Deployment



```
Final Predictions on Test Dataset:
[2. 0. 1. 1. 2. 0. 0. 1. 3. 2. 1. 0. 1. 2. 1. 2. 0. 0. 0. 2.]
```

The model demonstrates reasonable effectiveness for Class 0 but struggles significantly with Classes 1, 2, and particularly 3.

Feature Importance for Optimized Random Forest



Some key risk factors appear

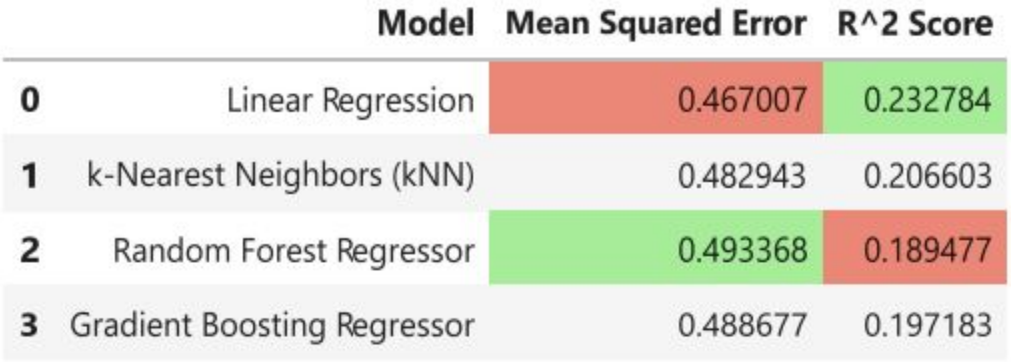
- **age of the child**
- **the level of sleep disturbance experienced**
- **hours per week of internet usage.**

Regression Models

Models tested: Linear Regression, kNN Regressor, Random Forest Regressor, Gradient Boosting Regressor.

Best model: **Linear Regression (R²: 0.233, MSE: 0.467)**.

Challenges in regression: Lower accuracy in predicting SII categories.



	Model	Mean Squared Error	R^2 Score
0	Linear Regression	0.467007	0.232784
1	k-Nearest Neighbors (kNN)	0.482943	0.206603
2	Random Forest Regressor	0.493368	0.189477
3	Gradient Boosting Regressor	0.488677	0.197183

Challenges and Solutions

Challenges:

- Missing values and incomplete features.
- Imbalanced class distribution.

Solutions:

- Imputation, SMOTE, and feature engineering.

Conclusion

- Best Model: **RANDOM FOREST**
- Classification performed better for SII prediction.
- Regression provided granular insights into PCIAT_Total but struggled with category accuracy.
- Some key risk factors appear to be the age of the child, the level of sleep disturbance experienced and - of course - hours per week of internet usage.
- **The model demonstrates reasonable effectiveness for Class 0 but struggles significantly with Classes 1, 2, and particularly 3.**

Errors in data [Performed Winsorization]

- A significant number of participants, especially for BMI and blood pressure, fall outside the expected normal ranges
- Most participants' heights and weights are within reasonable ranges, but many have BMIs outside the approximate normal range, suggesting that many participants may have disproportionate body proportions (or incorrect measurements?).
- Most of the **bioelectrical impedance analysis** data is highly skewed. The majority of participants have values at the extreme ends, with a few outliers that might be measurement errors. Some variables, like fat mass index and body fat percentage, even have implausibly negative values.