Predicting Internet Addiction in Children: A Comparative Analysis of Machine Learning Models

Ayodele Adeniyi

Contents

List of Figures

List of Tables

Abstact

The interest is in estimating how much a child spends on the Internet, and further, whether the child is addicted to the Internet. While Internet addiction itself is not a psychiatric disorder, parents need to know when their children have characteristics and behaviors associated with compulsive use of the Internet. The dataset to be used in exploring this relationship was collected by the Healthy Brain Network (HBN) from a clinical sample of about five thousand 5- to 22-year-olds who have undergone both clinical and research screenings. This report details the steps in estimating the total scores of the Parent-Child Internet Addiction Test, which assesses the likelihood of Internet addiction. We will begin by performing data cleaning and imputation for missing values, and Subsequently, we compare two methods of estimating the PCIAT scores—Elastic Net and Random Forest—and determine the method that has the least error. Our analysis revealed that Random Forest produced the lower Root Mean Squared Error (RMSE) value. The report further states the findings and recommendation of the study based on the analysis.

Section 1: Introduction

Internet addiction comes in various forms: net compulsions, cyber relationships, gaming, information seeking, and many more. Naturally, this is a concern, especially since according to Internet Live Stats (2016), 88.5% of the U.S. population used the internet, amounting to 286,942,362 users out of a total population of 324,118,787. The effects of this overuse are mainly felt by close relatives or colleagues of the user, as the addiction often puts a strain on their ability to fulfill both personal and professional obligations. University College London researchers reviewed several articles about adolescents diagnosed with Internet addiction, and the findings revealed that there were significant changes in brain networks of these adolescents. They observed both increased and decreased activity in the default mode network, which is usually active when the brain is at rest. Furthermore, they noticed a decline in the connectivity of the executive control network, which is responsible for thinking and decision-making. Given the importance of this subject, it has become essential to proactively study and inquire about the Internet use of adolescents and even adults to estimate when this Internet addiction may have developed for easier remedy.

The objective of the study is to compare two methods for predicting PCIAT scores, specifically we will be considering Elastic Net and Random Forest. We will use Elastic Net to predict the PCIAT score for internet users, we will also use Random Forest to perform the same task. We will then compare the performance of the two models based on the RMSE and recommend the better method based on the features. To achieve the goal of predicting PCIAT scores, the HBN dataset will be used. The original data set is sourced from Kaggle, and it contains 82 features and 3,960 observations. This includes: Age in years, Sex (gender of the participant), Weight in pounds, Sleep Disturbance Score, Hours of Using Computer/Internet in hours (Usage), Body Mass Index (BMI) in kilograms to the square of their height (in meters), among others. The original data set as well as details of the other features is available here.

Section 2: Data Cleaning and EDA

This section shows the exploratory data analysis of the dataset. Before proceeding to the different estimation methods on the PCIAT variable, we begin by checking for missing variables within the dataset, as this is the first step in our data cleaning process. The missing data check on the dataset revealed that about 35% of the data was missing. Section two details the analytical steps taken to address the missing data. Subsequently, we have a cleaned dataset with 57 features and 3,287 observations. This is the data to be used for this analysis.

From Table 1, it is noted that the age of the students ranged from 5 to 22, and the minimum PCIAT score is 0 with a maximum of 93, BMI ranged from a minimum of 8.522 to a maximum of 59.132, Hours of Using Computer/Internet in hours (Usage) averaged approximately 1.047 hours daily, the minimum Sleep Distance score is 17 to a maximum of 96, the minimum weight of the students is 31.80 pounds with a maximum of 315 pounds (see Table 1 for more details). Examining the data set for the presence of missing variables revealed that there are no missing variable, confirming that the is cleaned and imputed is data complete, and we can proceed to predict PCIAT scores using the different features.

Table 1: Descriptive Statistics of Features (continued below)

PCIAT	BMI	Usage	Age	Weight	SDS
Min. : 0.00	Min. : 8.522	Min. :0.000	Min. : 5.00	Min. : 31.8	Min. : 38.00
1st Qu.: 8.00	1st Qu.:15.722	1st Qu.:0.000	1st Qu.: 8.00	1st Qu.: 60.4	1st Qu.: 47.00
Median :24.00	Median :17.988	Median :1.000	Median :10.00	Median : 76.1	Median : 55.00
Mean :25.74	Mean :19.330	Mean :1.076	Mean :10.43	Mean : 90.1	Mean : 57.91
3rd Qu.:39.00	3rd Qu.:21.617	3rd Qu.:2.000	3rd Qu.:13.00	3rd Qu.:108.4	3rd Qu.: 66.00
Max. :93.00	Max. :59.132	Max. :3.000	Max. :22.00	Max. :315.0	Max. :100.00

The scatter plot of weight (x-axis) and total PCIAT score (y-axis) is shown in Figure 1.1 below. There is a positive correlation between weight and total PCIAT score, as the line of best fit slopes upward. This implies that children who weigh more tend to have higher PCIAT scores. The correlation is not very strong, as the points are distributed across the line.

The second scatter plot represents the relationship between Sleep Disturbance Score (x-axis) and PCIAT total score (y-axis), as shown in Figure 1.2 below. The relationship between the two variables has a positive slope, implying that children who experience higher sleep disturbances tend to have higher PCIAT scores. The data points are distributed across a wide range of Sleep Disturbance Scores (ranging from 25 to 100), with corresponding PCIAT scores ranging from 0 to 80.

The scatter plot in Figure 1.3 shows a positive relationship between Age and total PCIAT scores, suggesting that older children are more likely to exhibit high internet use. The median PCIAT score for males is higher than the median PCIAT score for females, as shown in Figure 1.4. Outliers are present in both categories; this is represented by dots above the whiskers. This suggests that in both groups, there are individuals who exhibit unusually high PCIAT scores compared to others of the same sex.

The median PCIAT score increases as the category of internet use increases, as shown in Figure 1.5. The "Less_than_1H_daily" category has the lowest median score, while the "More_than_3H" category has the highest median. Outliers are present in almost all categories, represented by dots beyond the whiskers.

The scatter plot of BMI (x-axis) and total PCIAT score (y-axis) is shown in Figure 1.6. It shows a positive correlation between BMI and total PCIAT score, as the line of best fit slopes upward. The data points are distributed across a wide range of BMI values (ranging from 10 to 60), with corresponding PCIAT scores ranging from 0 to 80. This implies that children with higher BMI tend to have higher PCIAT scores. The correlation is not very strong, as the points are distributed across the line.

Section 3: Chained Imputation - Multivariate Imputation by Chained Equations

Missing data is defined as values that are not observed. This section discusses a method of handling missing data. Specifically, the technique we will be using to handle missing data is called Multivariate Imputation by Chained Equations (MICE).

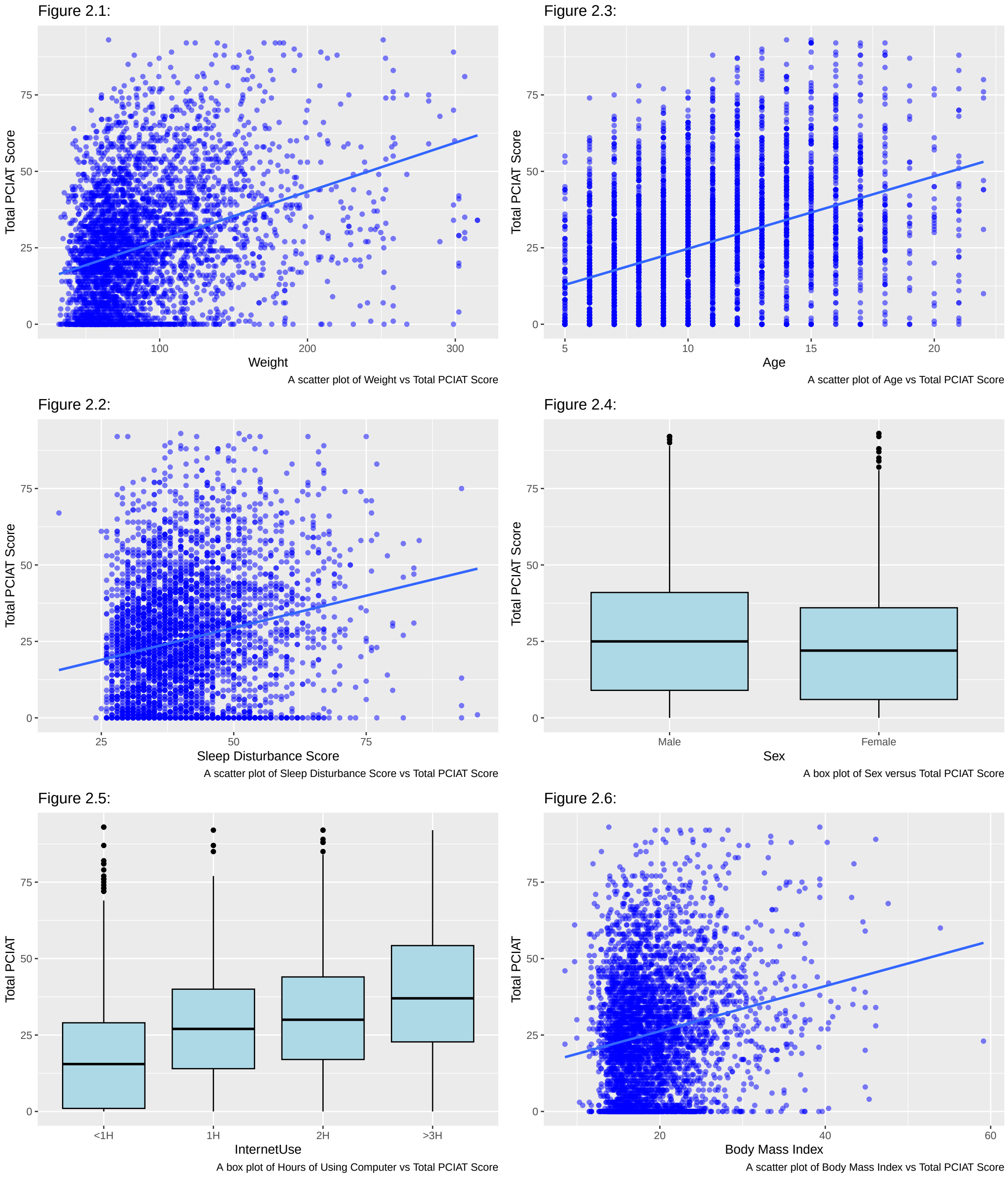


Figure 2.1:

A scatter plot of Weight vs Total PCIAT Score

Figure 2.3:

A scatter plot of Age vs Total PCIAT Score

Figure 2.2:

A scatter plot of Sleep Disturbance Score vs Total PCIAT Score

Figure 2.4:

A box plot of Sex versus Total PCIAT Score

Figure 2.5:

A box plot of Hours of Using Computer vs Total PCIAT Score

Figure 2.6:

A scatter plot of Body Mass Index vs Total PCIAT Score

Figure 1: Explatory Data Analytics of features vs PCIAT

Section 3.1: Introduction:

Missing data can be due to several factors, including, but not limited to, difficulty in obtaining that information during data collection, omission during data collection, and others. By default, R excludes missing data when creating models or plotting charts. However, allowing this would imply a statistical assumption called Missing Completely at Random. Simply stated, this means that missing data points are completely random, and thus excluding those missing observations creates no bias in the dataset. However, the assumption of Missingness Completely at Random is difficult to prove, as there is a possibility that the missing data may be related in some way, and thus excluding such data means excluding an important fraction of your data.

Ways of handling missing data include complete-case analysis. This is an approach that discards all rows with missing data. R performs a complete-case analysis while creating charts and visuals. The problem with this approach is the difficulty in proving missingness completely at random. Another challenge with using a complete-case analysis is that there may be significantly fewer observations to work with if a complete-case analysis is performed. In this instance, performing a complete-case analysis would reduce the number of observations to zero. This makes complete-case analysis impracticable for our analysis.

Other techniques for handling missingness include unconditional mean imputation. Mean imputation works by replacing the missing values with the mean of the observed values in that feature. While the benefit of this method is that it is easy to use, mean imputation can often distort the distribution of the variable, as it creates more variables that are zero standard deviations away from the mean. Hence, it ultimately underestimates the standard deviation of the variable.

Another estimation technique is regression imputation, which involves using a regression model to predict the missing values in our dataset. This method uses observed values of the variable with missing data as the dependent variable and the other variables as the independent variables to fit a regression model. It takes into account the relationship between variables in the dataset and can improve the accuracy of predictions. However, linear regression models are sensitive to outliers.

Section 3.2: Method:

This section discusses MICE as a method for imputation. MICE works by creating multiple imputations, as opposed to single imputations. The chained imputation approach is very flexible and can handle variables of varying types (e.g., continuous or binary). Azur, Stuart, Frangakis, and Leaf (2011) explained the steps for MICE as follows:

i. All missing values in the dataset are imputed using simple imputation methods, such as the mean. These imputations can be thought of as "placeholders."

ii. The "placeholder" mean imputations for one variable ("var") are set back to missing.

iii. In this step, the observed values from the variable "var" are regressed on the other variables in the imputation model, which may or may not include all the variables in the dataset. As a result, "var" is the dependent variable in a regression model, while all the other variables are independent variables. The regression models used here are based on the same assumptions as those used in linear, logistic, or Poisson regression models that do not involve missing data.

iv. Using the regression model, the missing values for "var" are replaced with predictions (imputations). Both the observed values and the imputed values will be used in regression models when "var" is subsequently used as an independent variable.

v. Steps 2–4 are repeated for each variable that has missing data. During one cycle, each variable is cycled, and its missing values are replaced with predictions based on regression analysis.

vi. Steps 2–4 are repeated for several cycles, with the imputations being updated at each cycle.

Section 3.3: Results:

This section discusses the results of the chained imputation. Following the application of MICE, we obtained a final dataset consisting of 57 features and 3,287 observations. This dataset will be used in our analysis. A limitation of MICE is that the method relies on the assumption that missing data is “Missing At Random” (MAR), meaning the probability of a value being missing depends only on observed variables, however, if this assumption is not met, the imputed data can be biased.

Section 4: Elastic Net

This section discusses the first method for predicting PCIAT scores based on 57 features and 2,287 observations. We will use Elastic Net to perform our first prediction of PCIAT scores.

Section 4.1: Introduction:

The choice of Elastic Net stems from the high correlation between the features in the dataset. Specifically, we noted a high correlation between height and age, and weight and age. This invalidates the use of linear regression. We can consider alternative methods like Lasso, Ridge, and Elastic Net regression. In Ridge regression, also known as L2 regularization, the method utilizes a penalty term to improve the ordinary least squares modeling. It is a shrinkage technique that shrinks coefficients towards zero, and a penalty term is added, this includes λ, a tuning parameter. This penalty comprises the tuning parameter multiplied by the squared sum of the coefficient values. On the other hand, Least Absolute Shrinkage and Selection Operator (Lasso) regression, also known as L1 regularization, adds a penalty term to the coefficients that is proportional to their absolute values. As a result, for high values of the tuning parameter λ, many coefficients are set to zero under Lasso. Thus, Lasso performs variable selection, which is not the case in Ridge regression. In data sets, Ridge performs shrinkage, whereas Lasso performs selection, even in instances when a shrinkage technique is better suited. Elastic Net emerged as a result of critiques of Lasso, whose variable selection can be too dependent on the data and thus unstable. The solution is to combine the penalties of Ridge regression and Lasso to get the best of both worlds. This is the technique we will use in this analysis.

Section 4.2: Method:

This section discusses Elastic Net regression as a method in predicting numerical variables. Elastic Net regression combines the strengths of Ridge and Lasso regression, involving both L1 and L2 penalties. Due to the Ridge regularization, the Elastic Net estimator can handle correlations between the predictors better than Lasso. Additionally, due to the L1 regularization, sparsity is achieved. For Elastic Net regression, we choose the estimates of $\hat{\beta}$ that minimizes the below:

$$\left[\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\hat{\beta}_j| + 1 - \alpha\sum_{j=1}^{p}\hat{\beta}_j^{\,2}\right)\right]$$

Where $\lambda \geq 0$ 0 and $1 \geq \alpha \geq 0$ are scalars. where α is the mixing parameter between ridge ($\alpha = 0$) and lasso ($\alpha = 1$).

We derive Elastic Net through the following steps:

i. Test sequence of alpha values from 0 to 1 and lambda values for a defined range.

ii. Perform 10-fold cross-validation on the features using Elastic Net regression for for each alpha value and each lambda value in the sequence.

iii. For each alpha value, store the optimal lambda and the RMSE/MSE.

iv. Determine which combination of alpha and lambda gives the best model by analyzing the RMSE/MSE results.

Specifically for our analysis, we consider λ values from 0 up to 25 with increments of 0.5, and α values within the range of 0 to 1, with increments of 0.01.

Section 4.3: Results:

This section discusses the results of the Elastic Net Regression method. The Root Mean Squared Error (RMSE) is a measure of the average magnitude of the error between the predicted values and the actual observed values. Lower RMSE values indicate a better fit of the model to the data. The mathematical expression for the RMSE is given below:

$$\text{RMSE} = \sqrt{\frac{1}{n_{test}} \sum_{i^*=1}^{n_{test}} (y_{i^*} - \hat{y}_{i^*})^2}$$

where y represents the true values of the response variable, and $\hat{y}$ denotes the predicted values.

Table 3: Optimal parameters for Elastic Net

	Alpha	Lambda	RMSE	MSE
17	0.16	0.5	16.44202	270.3401

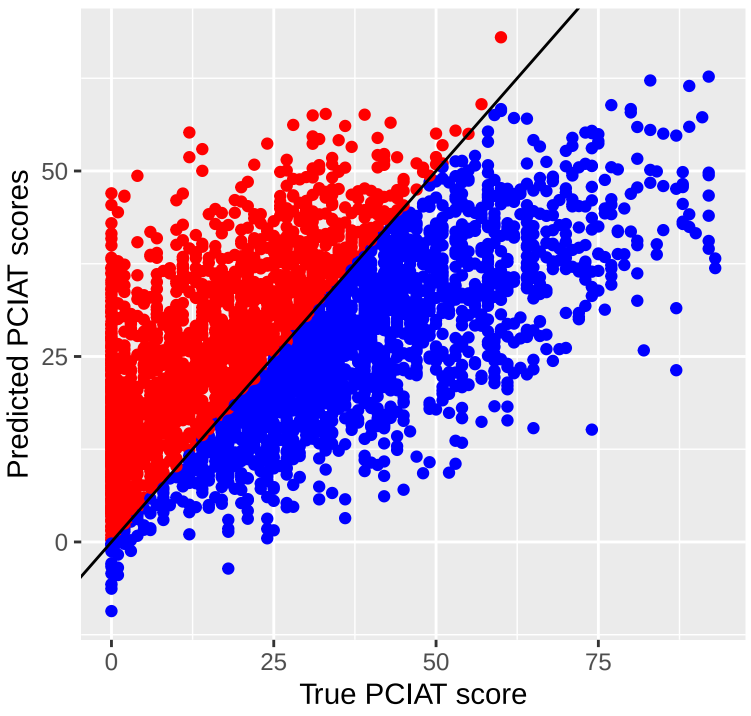
We considered λ values from 0 to 25, the optimal λ was 0.5. we considered α values from 0 to 1 and the optimal $\alpha = 0.16$, yielding a RMSE of 16.4420229 and MSE of 270.340117. This value means that, on average, our predictions for Total PCIAT score were off by approximately ± 16.4420229.We proceed to examine the relationship between the model predictions and the true PCIAT scores. This relationship can be viewed in a scatter plot with the true PCIAT score on the x-axis and the Elastic Net predicted PCIAT score on the y-axis. See Figure 2 below for more details. The red dots indicate where the model overestimated the PCIAT score, while blue points show where the model tends to underestimate scores, the points lie close to the regression line, showing that this is a pretty good prediction. It is important to note that the limitations of Elastic Net include the biased estimates of the coefficients that the regression method yields and the method is also computationally expensive.

Section 5: Random Forest

This section discusses the second method for predicting PCIAT scores based on the 57 features and 2,287 observations derived from the imputation. The final method is called Random Forest, a model that can handle both classification and regression problems. The choice of Random Forest is because the method can handle a wider range of feature types, it is less susceptible to outliers, and generally provides higher predictive accuracy when dealing with complex, non-linear relationships in data than Elastic Net.

Section 5.1: Introduction:

A decision tree is a type of flowchart that shows the pathway to a decision. "A decision tree starts at a single point (or 'node') which then branches (or 'splits') in two or more directions. Each branch offers different possible outcomes, incorporating a variety of decisions and chance events until a final outcome is achieved." A decision tree offers many benefits, which is why it is being considered. It interprets data in a highly visual way, works well with both numerical and non-numerical data, and can easily be combined with other decision-making techniques. However, some drawbacks of decision trees include the possibility of overfitting if the trees become too complex and bias when using an imbalanced dataset (i.e., where one class of data dominates another). A Random Forest combines the output from multiple decision trees. It creates a number of decision trees during the training process. Each tree is constructed based on a random subset of the dataset, and random subsets of features are used for each partition of the dataset. Randomness introduces variability



A scatter plot of True PCIAT Score vs Elastic Net Predicted Scores

Figure 2: A scatter plot of True PCIAT Score vs Elastic Net Predicted Scores

among individual trees, reducing the risk of overfitting and improving prediction accuracy. During prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or averaging (for regression tasks). The benefits of Random Forest include higher predictive accuracy compared to regression trees, resistance to overfitting, and better handling of large datasets.

Section 5.2: Method:

This section discusses the Random Forest algorithm. Random Forest algorithms have three main parameters: node size, the number of trees, and the number of features sampled. From these, the Random Forest classifier can be used to solve regression or classification problems. It is built on the idea of bootstrap aggregation, which is a method for resampling with replacement in order to reduce variance. Random Forest uses multiple trees to average (regression) or compute majority votes (classification) in the terminal leaf nodes when making a prediction. Built of the idea of decision trees, random forest models have resulted in significant improvements in prediction accuracy as compared to a single tree by growing 'n' number of trees; each tree in the training set is sampled randomly without replacement.

The number of features we would consider for the random forest method is 8, while the number of trees to be grown is 1000.

Section 5.3: Results:

This section discusses the results of the Random Forest algorithm. Based on the random forest method using the 1000 trees and 8 variables randomly sampled as candidates at each split, we obtained an RMSE of 16.9315969. This value means that, on average, our predictions for Total PCIAT score were off by approximately ± 16.9315969. We proceed to examine the relationship between Random Forest predictions and the true PCIAT scores. This relationship can be viewed in a scatter plot with the true PCIAT score on the x-axis and the Random Forest predicted PCIAT score on the y-axis. See Figure 3 below for more details.

The red dots indicate where the model overestimated the PCIAT score, red points are prevalent in the left area of the plot, indicating that the model tends to overestimate in this range of PCIAT values (i.e., from 0 to 40), while blue points are more prevalent in the right part of the plot, showing that the model tends to underestimate scores when PCIAT values are greater than around 50, given the range of our PCIAT scores, this is also a very good prediction and most points are very close to the line with little dispersion observed. However, this method has the disadvantage of being computationally expensive.

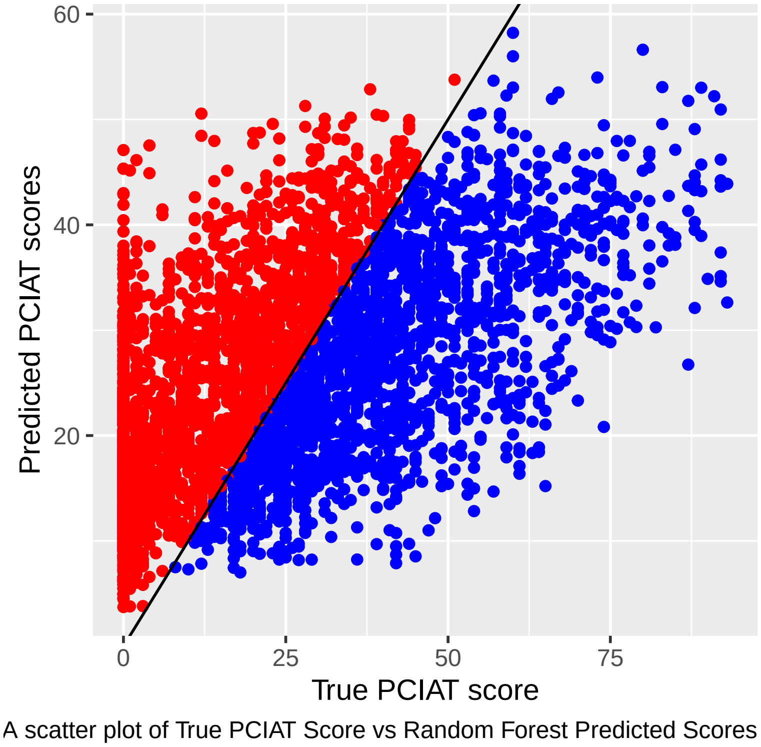


Figure 3: A scatter plot of True PCIAT Score vs Random Forest Predicted Scores

Conclusion

The objective of this study is to determine the most effective predictive model for estimating scores on the Parent-Child Internet Addiction Test (PCIAT) based on features. We considered two models: Elastic Net regression and Random Forest. Elastic Net combines the advantages of both Ridge and Lasso regression, yielding a lower RMSE of 16.4420229, while Random Forest, with its ensemble of multiple decision trees working to improve overall prediction accuracy, yields an RMSE of 16.9315969. In conclusion, the analysis revealed that Elastic Net regression outperforms Random Forest in predicting PCIAT scores, achieving the lower RMSE of 16.4420229. The results indicate that, on average, the predicted Parent-Child Internet Addiction Test (PCIAT) scores are approximately 16.4420229 away from the true values. I would recommend the use of Elastic Net for the prediction of PCIAT scores, however, it is important to note that the limitations of Elastic Net include the biased estimates of the coefficients that the regression method yields and the method is also computationally expensive.