Team 6- Child Mind

Jacob Hall, Caleb Smith, Jaden Johnson, Daniel Rodriguez

1. Problem Statement and Background (Jacob – 25%, Caleb – 25%, Jaden – 25%, Daniel – 25%)

Can you predict the level of problematic internet usage exhibited by children and adolescents, based on their physical activity and internet usage behavior?

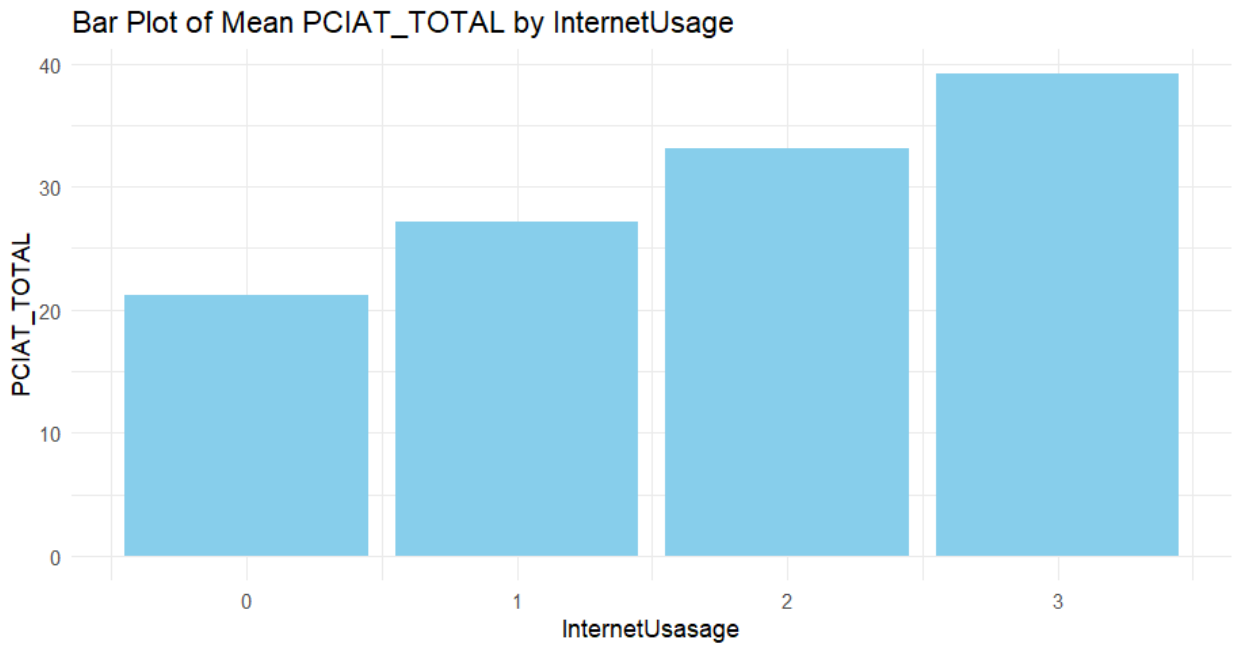
The data is from The Healthy Brain Network (HBN) dataset which is a clinical sample of about five-thousand 5–22-year-olds who have undergone both clinical and research screenings. The data has two elements which are being used in this project: physical activity data (wrist-worn accelerometer data, fitness assessments and questionnaires) and internet usage behavior data. The prediction will be represented by the Severity Impairment Index (0-none,1- Mild,2-Moderate,3-Severe). Identifying these patterns can help trigger interventions to encourage healthier digital habits.

2. Data and Exploratory Analysis (Jacob – 25%, Caleb – 25%, Jaden – 25%, Daniel – 25%)

The data set of train.csv has 82 unique features(columns) and 3960 unique data points(rows). These features are categorical and numerical values to describe the data points' physical activities, physical measures, sleep quality, and their internet activities. For example, there are features such as PreInt_EduHx-computerinternet_hoursday, which is hours of using computer/internet, and BIA-BIA_Fat, which is their body fat percentage. 79 of the features had NA's that we had to clean, and some rows had outliers that needed cleaning/removing as well. The only ones that had no issues was there id, age, and sex.

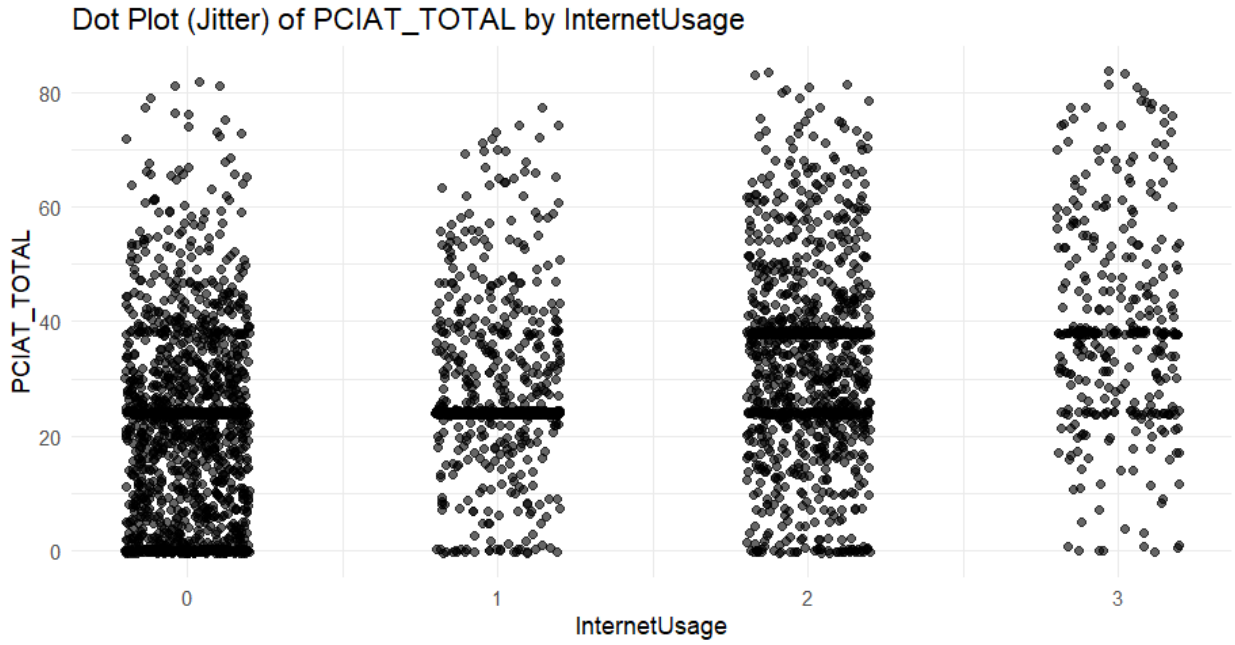
We used tools in the tidyverse library to extract, clean, and generate data. Using these libraries, we read in the data using read.csv functions and cleaned it by first removing all the seasons from the data because we had concluded that to predict the sii score based

on physical activity, we would not need to know what seasons the data was collected in. Then we removed all the outliers for the data because some numbers did not make sense to keep in for example: The value for one of the data points BIA-BIA_BMR (Basal Metabolic Rate) was 83152.2, which was almost 70x more than the average for that feature. After removing the outliers, we averaged out all the features based on age groups and input the averages for the features with missing values. The age groups we used are the ones that the competition provided which are: Adolescents (13 to 22) and children (5 to 12). Now with this clean data we should be able to start making some of our own predictions and see what models we should use too.



This graph shows PCIAT_TOTAL by PreInt_EduHx-computerinternet_hoursday. PreInt_EduHx-computerinternet_hoursday is hours of using computer/internet a day based on a categorical scale of 0=Less than 1h/day, 1=Around 1h/day, 2=Around 2hs/day, 3=More than 3hs/day. PCIAT_TOTAL is just the SII score but in a more well-rounded categorical way: Severity Impairment Index: 0-30=None; 31-49=Mild; 50-79=Moderate; 80-100=Severe. The graph shown above shows a clear correlation between the two variables.

As you spend more time on your computer, you're more than likely to have problematic internet usage and need to go out more.



This graph shows PCIAT_TOTAL by PreInt_EduHx-computerinternet_hoursday. PreInt_EduHx-computerinternet_hoursday is hours of using computer/internet a day based on a categorical scale of 0=Less than 1h/day, 1=Around 1h/day, 2=Around 2hs/day, 3=More than 3hs/day. PCIAT_TOTAL is just the SII score but in a more well-rounded categorical way: Severity Impairment Index: 0-30=None; 31-49=Mild; 50-79=Moderate; 80-100=Severe. The graph shown above is just another way of observing this relationship between the two features.

While there does appear to be a relationship between quantity of internet usage and a subject's score on the Parent-Child Internet Addiction Test, the purpose of this problem is to be able to predict a subject's SII (derived from the Parent-Child Internet Addiction Test score) based on both internet usage measures *and* physical activity measures. Going forward, we will need to do further exploration to find physical activity-related attributes that appear to have a strong relationship with a subject's SII/PCIAT Score and find the best

attributes to use as predictors. Because not all SII scores are labeled in the training data, we may also need to use some unsupervised learning techniques to cluster together subjects with similar attributes.

3. Methods (Jacob – 25%, Caleb – 25%, Jaden – 25%, Daniel – 25%)

Using the cleaned data mentioned in the previous section the first model we tried was linear regression made by Jacob hoping that it would be able to predict sii using some values and we tried it with a few values, but all seemed to give us the same prediction power, so we will just give one example we did which is the BIA.BIA_ICW (Intracellular Water which means a lot of it would equal large muscle mass). It had very low predicting power and this was the case for most when using this model, so this model was demined a failure.

After linear regression, Jaden tried a step further with a multivariate polynomial regression model of degree 2. For the model, we initially chose to limit the features down to 7 to prevent any overfitting. Furthermore, PCIAT values were not included in the list and the highest correlating features were selected. Using these 7 produced slightly better prediction power than the linear model. However, this is still very low and was deemed a failure. To test the model further, we also tested the code by changing the degree to 3, which produced a better result but still a failure and changed the number of features to 10, which produced relatively the same results meaning failure once again. Neither of the changes caused any significant increase in predicting power.

Within the data we had categorical data represented numerically, we converted the categorical variables to factors levels with meaningful labels. For example, the feature Basic_Demos.Sex was represented as 0 for male and 1 for female which were converted for a more interpretable dataset. We used the data dictionary provided for use from the Kaggle competition to accurately convert these categorial features.

The next algorithm we tried was Random Forest Classification by Daniel. The reason was because it is able to handle complex relationships, it is robust to overfitting and provides a metric for being able to see which features are important. At first when we ran the model, we noticed we did not remove PCIAT, which is Parent-Child Internet Addiction Test. The PCIAT is the metric used to help categorize the Severity Impairment Index (sii), which is what we are trying to predict. Thanks to the visualization tool vip() we caught that mistake. We went ahead in removing any feature that involved PCIAT.

Once we removed PCIAT features we ran the model again. The model was configured with 500 trees, with a minimum node size of 8, with the split variables at 9. Our model strategy involved splitting the data into training at 70% and testing at 30%. The model was assessed by examining a confusion matrix and the out-of-bag error rate. The results revealed that for some classification categories the error rate was high, within the confusion matrix this error was caused by an overfitting problem even though the Random Forest Classification was supposed to be a robust model. The overfitting came from the data having more data in the categories of sii being 0-None and 2-Moderate in fact over half the data was classified into those categories.

Seeing this issue we immediately got to work on ways we could fix this and the solution we decided on was injecting dummy data into the clean data to add more to the categories 1- Mild and 3-Severe. To do this we made a function that generated synthetic data for new participants, ensuring it adhered to the existing data structure. This function simulated data for various health and behavioral metrics. The data for each participant was created using random sampling from appropriate ranges or distributions, ensuring variability but maintaining biological plausibility. So, the function for example wouldn't input 300 for height or 1000 for weight. To address the problem, 2000 new synthetic participants were generated with adjusted biases. After generating the new data, it was appended to the existing dataset to create a larger and more balanced data frame. This

augmentation step was crucial in improving the representativeness of the dataset, particularly for the 1-Mild and 3-Severe categories, which had fewer instances initially.

After fixing the data set, Jacob wrote a new Random Forest Classifications model we once again converted the sii to factor levels with meaningful labels. which were converted for a more interpretable dataset. Then we cleaned up any NA's that could have been left over via the cleaning of the data. Then we converted categorical features into binary dummy variables to allow the model to properly interpret the data. Then just so we don't have the issue with the oversampling of the 0-None and 2-Moderate categories we generated some more data to help the model because 1-Mild and 3-Severe were dominating the other categories. The model was still configured with 500 trees, with a minimum node size of 8, with the split variables at 9, and we also changed the engine in this model to ranger. We also kept the strategy involving splitting the data into training at 70% and testing at 30%. We then visualized everything using a confusion matrix and saw that the predicting capabilities of this model were a lot stronger than the last and deemed this one a success.

As an alternative attempt, we tried a different method of data cleaning that involved imputing missing sii scores based on clustering rather than mean (the description for the Kaggle competition suggests using unsupervised learning techniques). We used k-means clustering based on all the numerical and categorical integer attributes (minus a few that were only applicable to certain age groups and therefore missing much data). Four clusters were used, as suggested by the elbow method from visualizing the within-cluster-sum-of-squares for different numbers of clusters, and also because there were four different possible values for the sii score. (Note: the missing values in the predictor columns were still imputed using the mean for two age groups: 5-12 and 13-22.) We then filled in the missing sii scores based on the mode of the sii score for each cluster. This data was then used in the second version of our Random Forest Classification model to see if it would perform better.

4. Tools (Jacob – 25%, Caleb – 25%, Jaden – 25%, Daniel – 25%)

For this part of our project, we utilized the R programming language with the RStudio Integrated Development Environment for reading, cleaning, visualizing, and analyzing data. We specifically used tools from external libraries such as tidyverse, tidymodels, themis, and ranger, and utilized various machine learning models to try to find a relationship between potential predictor variables and the sii score that we were trying to predict. The tools for the linear regression were simple it was just ggplot library for visualizing and base R for lm() which is the linear regression model.

The tools used for the multivariate polynomial regression model of degree 2 were ggplot library for visualization, base R for lm() which as the polynomial regression model and cor() for finding r^2, and metrics library for producing mean squared error (MSE).

The tools used for the first Random Forest Classification was R studio, because it provides a good UI to manipulate and visualize data. We utilized the Tidyverse ecosystem, with packages like dplyr for data manipulation and tidymodels for creating machine learning workflow. We also used the package vip() from RandomForest to allow us to visualize valuable feature importance metrics. The data prepossessing tools used were step_impute_mean() to handle missing numeric values, step_novel() to manage unseen factor levels and step_dummy() to convert categorical to numeric. For model evaluation of the Random Forest Classification, we used conf_mat() which helps generate a confusion matrix and autoplot() to visualize the confusion matrix in a heat map.

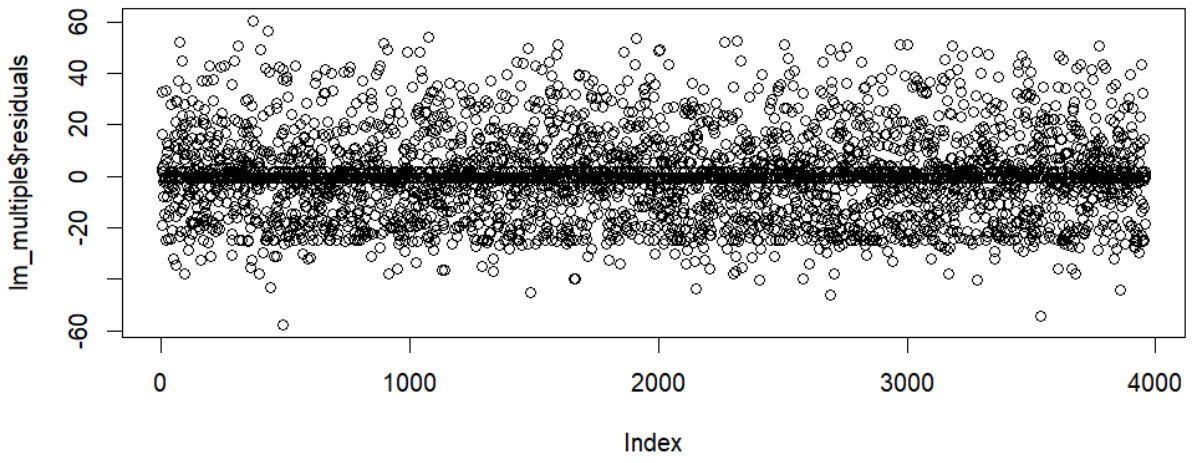
The tools for the second Random Forest were the exact same as the previous one that had the overfitting problem with a few exceptions. Those exceptions are the libraries' ranger and themis. The library ranger is an engine used for fitting for Random Forest Classification. The reason behind using this one is because it is an efficient implementation of Random Forest for both classification and regression tasks. It is fast and scalable, capable of handling large datasets. The ranger algorithm is particularly useful when you have many variables and interactions between them, as in this case.

Using this made the model faster to process with the previous model results took about 20-40 mins because how big the data set this model literally cut that time in half or possibly more. The library themis was used for SMOTE (Synthetic Minority Over-sampling Technique) was used to balance the class distribution in the training data by generating synthetic samples for the minority classes. The function from the library that does this is step_smote(sii) which was used in the script to apply SMOTE to the sii variables during training. Then as stated before we used tidymodels for visualizing.

For the k means clustering, we used the kmeans function from base R. We also used the select function from dplyr to take out columns that would not be used for the clustering. TutorialsPoint and ChatGPT helped us write a function to get the mode of a column in a data frame.

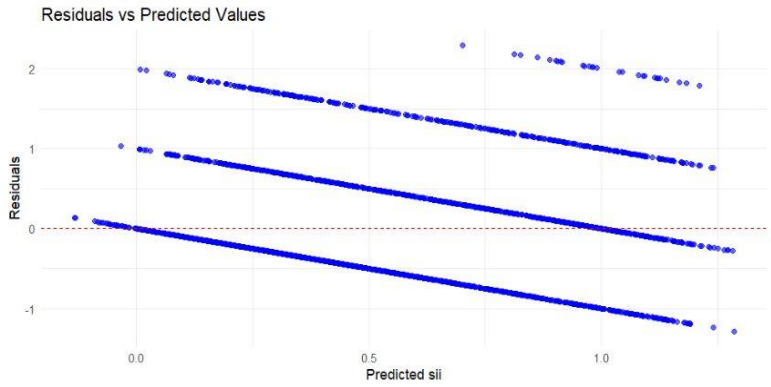
5. Results (Jacob – 25%, Caleb – 25%, Jaden – 25%, Daniel – 25%)

As previously stated in the section the first model used was the simplest one, we could think of linear regression. You could think of this one as our baseline model where we went through and checked variables to see if they had a strong relationship with the PCIAT.PCIAT_Total (Parent-Child Internet Addiction Test Total Score). For example, one of the usages would be BIA.BIA_ICW (Intracellular Water which means a lot of it would equal large muscle mass). The thought process behind this is that the bigger this value is the more muscle mass this person/child would have meant that they possibly might now use the internet that much if any which would mean that they would score low on the sii and true for the opposite as well. So, using the linear regression model we got an R-squared score of .1715036 meaning that it showed no too little relation and to further visualize this here is a graph.



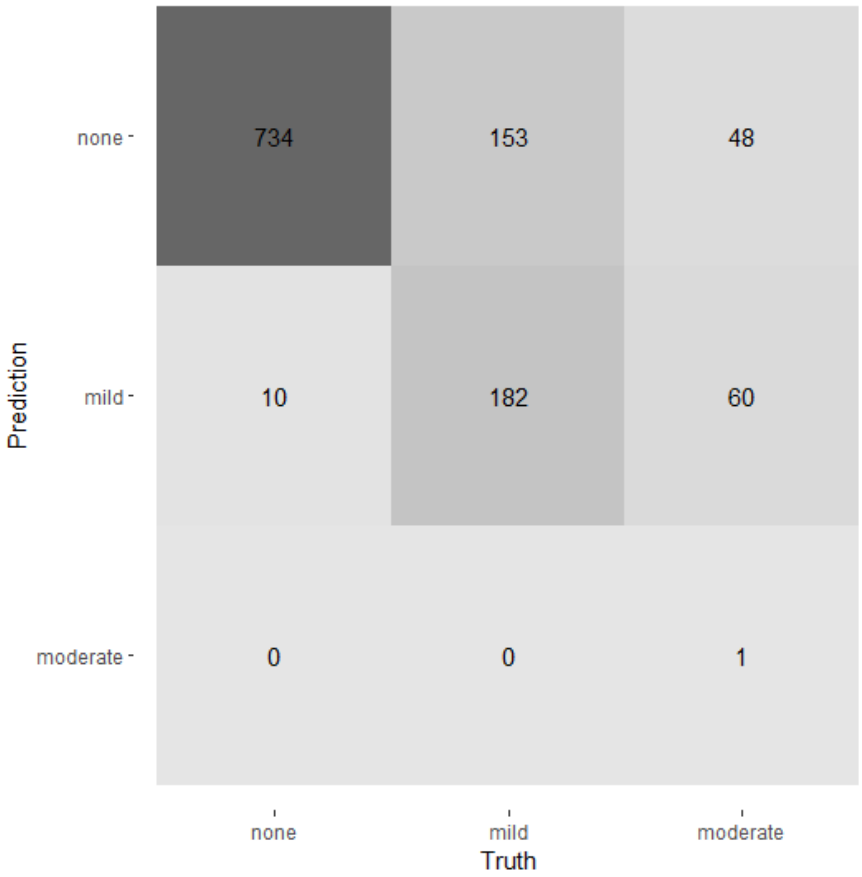
What this graph is showing is the residuals of the linear regression. How this helps show little to no relation is how the dots are scattered very far away from zero. If there was a relation all the dots would be scattered very close to zero or on zero showing a relation however this graph shows almost just a huge mess where not many of the dots fit this format. So, our simplest metric of finding a relation between physical features and sii score failed.

The next approach used was polynomial regression. We thought that this method may be slightly more accurate than the linear regression model since it can use multiple features to predict the sii. Various forms of the model with different degrees and non-PCIAT features were used. used. First, we used a 2-degree model with the top 7 highest correlating features to prevent over-fitting. It had an MSE (mean squared error) of 0.3402783 and an R^2 score of 0.2273093. This is slightly more accurate than the linear regression but still lacking. Next, we tried a higher degree of 3. It generated an MSE of 0.3376076 and an R^2 of 0.2333739. This is barely more accurate than the last model. Lastly, we tried the model of degree 2 with the top 10 features instead. This produced an MSE of 0.3394866 and an R^2 of 0.229107. Since the R^2 was still relatively low for all the models, it fails. All models had a similar graph like the one below.

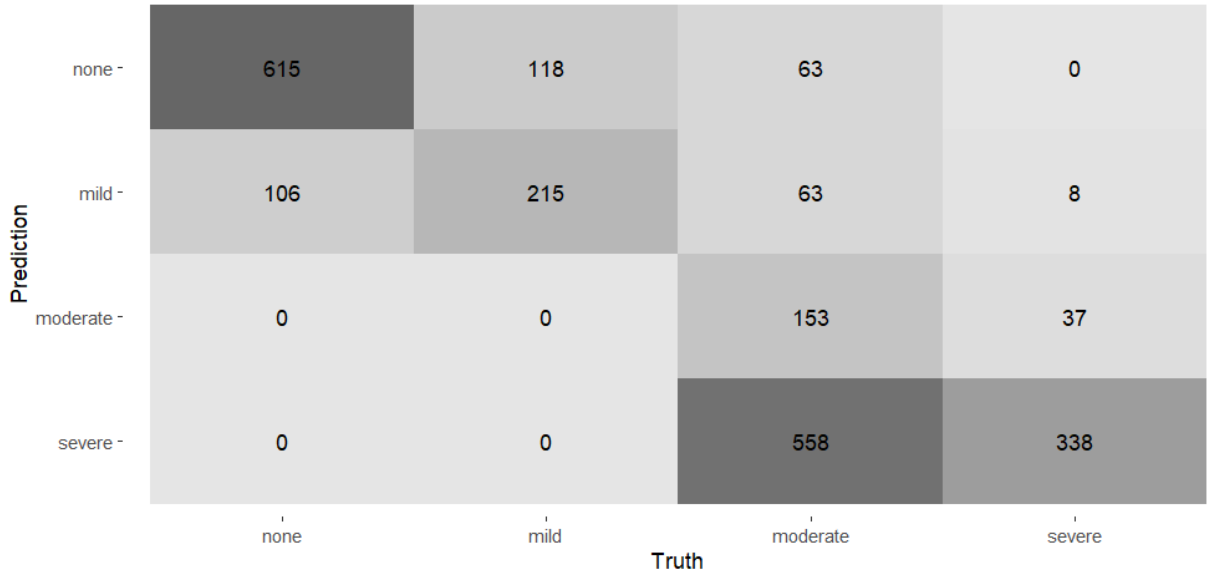


In the graph, the residual errors have a pattern which implies that the model is not a good fit. Since the R^2 was still relatively low for all the models and the graph showed poor results, this model fails.

After trying linear and polynomial regression we went ahead and tried random forest classification. Random forest is well suited to handling data sets with multiple features, as well as it is robust in handling overfitting. When running the model we encountered an error. This error was due to having minimal data that contained the severe class for our predicted feature sii (severity impairment index). To solve this, we went ahead and just removed the severe class from the prediction. The metric we used to measure this model was OOB (Out-of-Bag) Error, which had an estimated error rate of 23.92%. Another metric used was a confusion matrix and class errors. None had a class error of 0.01759531, mild's class error was 0.45989975, moderate class error was 0.98884758. This iteration of this model fails since there is a high-class error for moderate and we were not able to predict the severe class of sii.



The last and final approach we had was the idea of a different style of RandomForest that would generate more data of the values that where under sampled in the data making the training even better because it would have more of every category to train on doing this did help some and we do think of this as a success but there is some down side to this.



As shown in this heat map which we use to rate our model's performance to see how it did It has a fairly good idea if you are in the sii category none or severe. Where the problem starts to appear is that even when injecting dummy data, we can only simulate to the best of our abilities what a person in a said category might be like. Meaning when generating data sometimes it can have a harder time ranking the moderates in the right category. However, we still think this is better than not generating data because as you saw in the previous graph of the RandomForest that did not use generated data, it had trouble identifying any moderate or severe. So, when asking the question can you predict the level of problematic internet usage exhibited by children and adolescents, based on their physical activity and internet usage behavior? Using this model with some more data gathered we think it is possible to predict problematic internet usage exhibited by children and adolescents because it can be trained on the data points that we have a better amount of. It's the ones that we had to create that have some defects that it gets tripped up on.

As mentioned in section 3, the second Random Forest Classifier was attempted again with a dataset that was cleaned with a different technique (missing sii scores imputed based on k-means clustering). The idea was that using the mean technique to impute missing sii scores was causing the desired relationship to become obscured by having such a large portion of the sii scores to just be the average of the others. After running the Classifier with this modified data, we got these results:

Unfortunately, as evidenced by the heatmap confusion matrix, the Random Forest Classifier appeared to perform *worse* on predicting most classes with this new method of imputing the missing sii scores. However, it did perform slightly better on predicting the Moderate class, so perhaps this idea could be adapted for a better model.

6. Summary and Conclusions (Jacob – 25%, Caleb – 25%, Jaden – 25%, Daniel – 25%)

So, looking back at all the models we used and all the ways we manipulated the data for cleaning and for different results we can say with confidence that with more data points spread out into different sii categories for training the model you could use the last Random Forest model discussed in the results section as a model for predicting problematic internet usage for adolescents and children. With this model it has the capacity to handle robust and large data sets and will even help when the gap between respiration in categories is small. Although the model failed to accurately predict the

Moderate category, its success with the other 3 categories indicates that the relationship can be predicted with a more complete dataset.

7. Appendix

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

Works Cited

"The author(s) acknowledge the utilization of ChatGPT, a language model developed by OpenAI, in the preparation of this assignment. ChatGPT was employed in the following manner(s) within this assignment: "Fixing broken R code, writing an R script to replace NA values with average, writing an R script to get the mode of a column in a data frame while ignoring NA values, figuring out string interpolation in R, explaining an R function that was written to remove outliers, writing R scripts, and explaining errors."