https://doi.org/10.29013/ELBLS-23-1-25-31

Xiangbo Guo,

BUILDING A PREDICTIVE MODEL OF ADHD AMONG CHILDREN

Abstract. Attention Deficit/Hyperactivity Disorder (ADHD) is one of the most common neurodevelopmental disorders of childhood. According to the Centers of Disease Control and Prevention (CDC), the estimated number of children ever diagnosed with ADHD nationalwide is 6.1 million (9.4%). Among those children, 6 in 10 with ADHD had at least one other mental, emotional, or behavioral disorder that may have long-lasting impacts on their development.

In this research, we investigated possible risk factors related to development of ADHD among children and identified the most significant positive and negative factors through logistic regression. We used the 2020 National Survey of Children's Health survey data containing 42.777 complete data samples with features ranging from demographic information to the child's family condition. The response variable is whether a child has ever been diagnosed with ADHD.

After processing the dataset, we built a logistic regression model to predict whether a child will develop ADHD. By investigating the logistic regression coefficients, we found that parents' physical and mental health, the family's financial ability to cover basic living expenses, and whether the parents are divorced are all risk factors. The logistic regression model has achieved an AUROC score of 0.73, with 0.67 true positive rate (TPR) and 0.324 false positive rate (FPR). This predictive model is helpful for healthcare professionals to identify and reduce the risk for the children that are prone to the development of ADHD.

Keywords: ADHD, logistic regression, model, ROC, children, risk.

1. Introduction

Attention Deficit/Hyperactivity Disorder, or ADHD, is usually first diagnosed in childhood and often lasts into adulthood. Children with ADHD may have trouble paying attention, controlling behaviors, and sometimes appear to be reckless. It is normal for children to have trouble in focusing and behaving at one time or another, including failing to give close attention to details, making careless mistakes in schoolwork, at work, or with other activities, and having trouble organizing tasks and activities. In addition, children diagnosed with ADHD also appear to be hyperactive, with typical sympotoms like being overly talkive and easy to be annoyed.

The estimated number of children ever diagnosed with ADHD, according to a national 2016 parent survey, is 6.1 million (9.4%). This number includes: 388,000 children aged 2–5 years; 4 million children aged 6–11 years; 3 million children aged 12––17 years. Boys are more likely to be diagnosed with ADHD than girls (12.9% compared to 5.6%) [1]. According to a national 2016 parental survey, 6 in 10 children with ADHD had at least one other mental, emotional, or behavioral disorder [2].

ADHD is believed to be cuased collectively by multiple factors, including genetic, such as familial inheritance, food additives/diet, lead contamination, cigarette and alcohol exposure, maternal smoking during pregnancy, and low birth weight [3]. The sympotoms of ADHD can be alleviated through ways like mental counseling together with stimulant or nonstimulant medications.

Given the major impact of ADHD on patients' daily life, it is of great importance for healthcare professional to identify children that are at high risk for developing ADHD and help address problems at an early stage. To fulfill this task, this report discussed the machine learning techniques that can be applied to build predictive models on whether a child will develop ADHD. Specifically, we pre-processed the dataset, built a logistic regression model, and investigated factors most related to the development of ADHD. We also measured the model performance using various validation techniques and analyzed the model coefficients to find the variables that contribute most to our predicted results.

2. Method

2.1 Data

We used 2020 National Survey of Children's Health survey data for this study. The National Survey of Children's Health (NSCH) is conducted by the U. S. Census Bureau for the U. S. Department of Health and Human Services' (HHS) Health Resources and Services Administration's (HRSA) Maternal and Child Health Bureau (MCHB). It is designed to provide national and state-level information about the physical and emotional health and wellbeing of children under the age of 18 in the United States, their families and their communities, as well as information about the prevalence and impact of children with special health care needs. The 2020 NSCH data contains 42.777 complete data samples. We used the following variables as independent variables.

Table 1.– Features used for analysis

Variable	Description	Comments
HHCOUNT	How many people are living or staying at this address?	
A1_BORN	Where were you born?	1: In the U.S., 2: Outside the U.S.
A1_GRADE	What is the highest grade or level of school you have completed?	Higher value indicates higher education
A1_MARITAL	What is your marital status?	
A1_AGE	What is your age?	
A1_PHYSHEALTH	In general, how is your physical health?	Higher value indicates worse physical health
A1_MENTHEALTH	In general, how is your mental or emotional health?	Higher value indicates worse mental health
SC_SEX	What is this child's sex?	1: Male, 2: Female
SC_RACE_R	What is this child's race?	
AGEPOS4	Birth order of this child.	
SC_HISPANIC_R	Is this child of Hispanic, Latino, or Spanish origin?	1: Yes, 2: No
BIRTHWT_L	Low birth weight (< 2500g).	1: Yes, 2: No
ACE1	SINCE THIS CHILD WAS BORN, how often has it been very hard to cover the basics, like food or housing, on your family's income?	Higher value indicates higher frequency
ACE3	Has this child EVER experienced any of the following? Parent or guardian divorced or separated	1: Yes, 2: No
ACE4	has this child EVER experienced any of the following? Parent or guardian died	1: Yes, 2: No

The dependent variable is a binary feature coded as "K2Q31A," which indicates whether the child has ever been diagnosed with Attention Deficit Disorder (ADD) or Attention Deficit/Hyperactivity Disorder (ADHD).

2.2 Exploratory Analysis

A correlation graph is a primitive yet straightforward representation of the cells of a matrix of correlations. The idea is to display the pattern of correlations in terms of their signs and magnitudes by using visual thinning and correlation-based variable ordering. Moreover, the matrix cells can be shaded or colored to show the correlation value. The positive correlations are shown in red, while the negative correlations are shown in blue; the darker the hue, the greater the magnitude of the correlation.

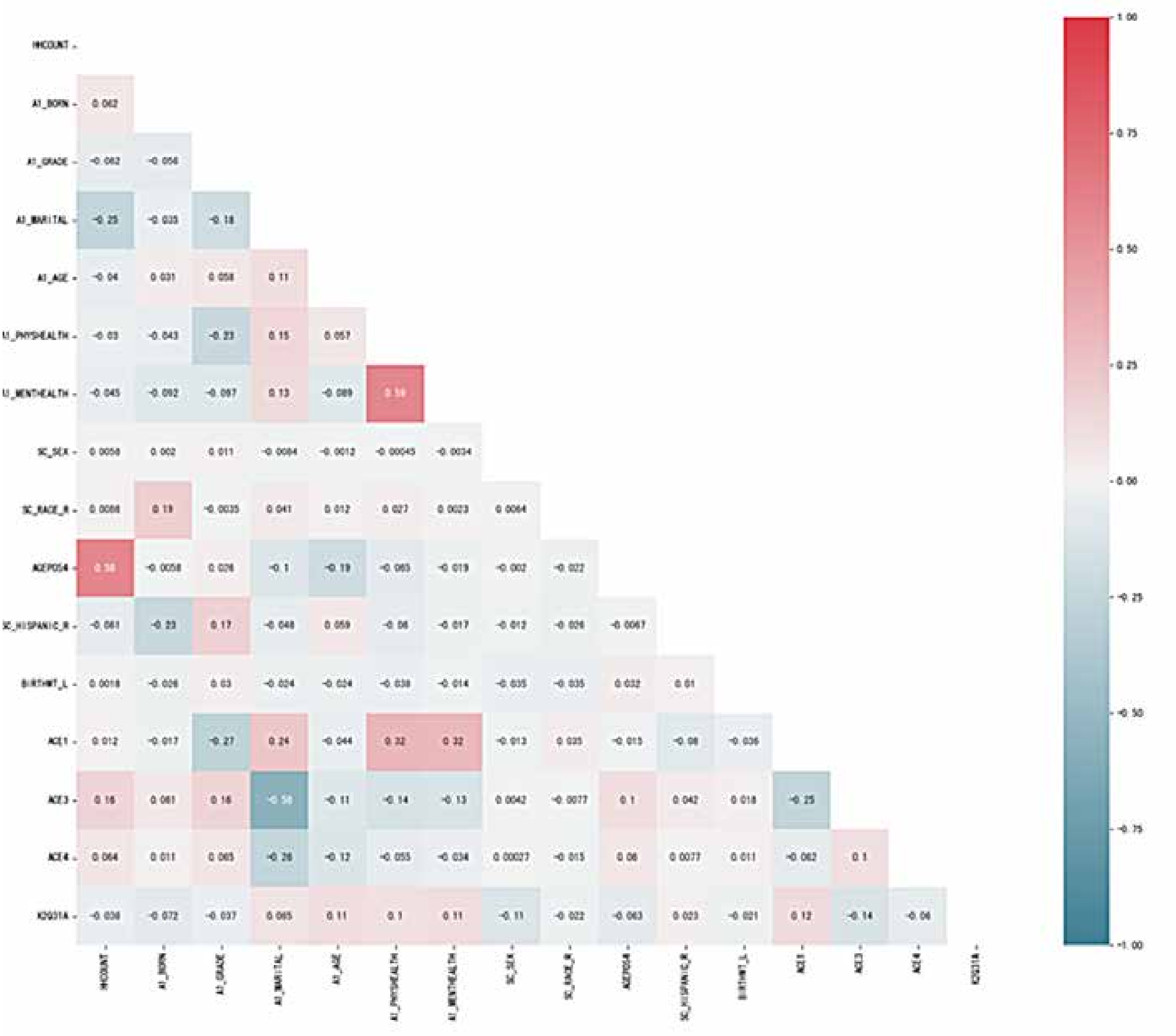


Figure 1. Correlation among variables

The graph above shows that the dependent variable (has ADHD) has the highest positive correlation with ACE1 (hard to cover basic living expenses), while having the highest negative correlation with ACE3 (parents or guardians not divorced).

In addition, the correlation graph also provides valuable information regarding the relationship among features. For example, the correlation between A1_PHYSHEALTH (parents' physical health condition) and A1_MENTHEALTH (parents' mental health condition) is 0.59, indicating that the two variables are significantly positively correlated and generally parents with worse physical health condition may also have worse mental health condition.

2.3 Statistical Method

2.3.1 Pre-processing

The data set is pre-processed in this step to improve both the training speed and accuracy. As the dataset is complete and does not contain any missing values, we did not employ any imputation technique here. In addition, as different features usually have remarkedly different value ranges, we applied the feature standardization technique to transform different features into comparable scales. This measure ensures that different features weigh equally in the training process. For each feature, its mean value and standard deviation are first computed as *avg(x)* and *std(x)*. Then each data point x with respect to that feature is replaced by y_i calculated as:

$$y_i = \frac{x - avg(x)}{std(x)} .$$

Finally, the dataset is partitioned into two datasets for training and test purposes: the training dataset (70%) for model development and the test dataset (30%) for model test and validation.

2.3.2 Logistic Regression

Logistic regression models were used to calculate the predicted risk. Logistic regression is a part of a category of statistical models called generalized linear models, and it allows one to predict a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a combination of these. Typically, the dependent variable is dichotomous, and the independent variables are either categorical or continuous.

The logistic regression model can be expressed with the formula:

$$\ln\left(\frac{y}{1-y}\right) = w_0 + w_1 x_1 + \ldots + w_m x_m$$

In the logistic regression, *y* is the probability of the sample classified as the positive class, and each feature x_i has its specific weight w_i, where w_0 is the intercept while w_1 through w_m are the coefficients of the independent variables.

Our task is to find a set of parameters $w_0, \ldots, w_m$ such that the loss function between the output *y* and the actual values *u*

$$l(y,u) = \| y - u \|_2^2$$

is minimized.

In addition, we applied elastic-net regularization to constrain model complexity and prevent model over-fitting problems with L-1 ratio equaling 0.5.

2.3.3 Model Validation

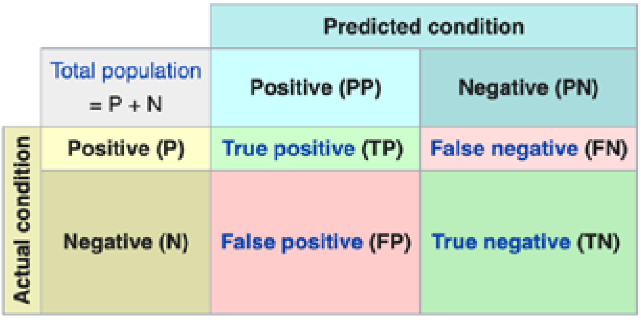
Consider a two-class prediction problem, where the outcomes are labeled either as positive or negative. There are four possible outcomes from a binary classifier. If the outcome from a prediction is positive and the actual value is also positive, then it is called a true positive (TP); however, if the actual value is negative, then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are negative, and false negative (FN) is when the prediction outcome is negative while the actual value is positive. In this way, the true positive rate (TPR) can be calculated as follows:

$$TPR = \frac{TP}{TP + FN}$$

And the false positive rate (FPR) can be calculated as:

$$FPR = \frac{FP}{TN + FP}$$

A confusion matrix is a table that allows visualization of the performance of an algorithm. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class. An example of the confusion matrix and the meaning of each cell within the table can be found in the graph below. Typically, the confusion matrix of a good predictive model has high true positive and true negative rates.



		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Figure 2. Confusion matrix example

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings [4].

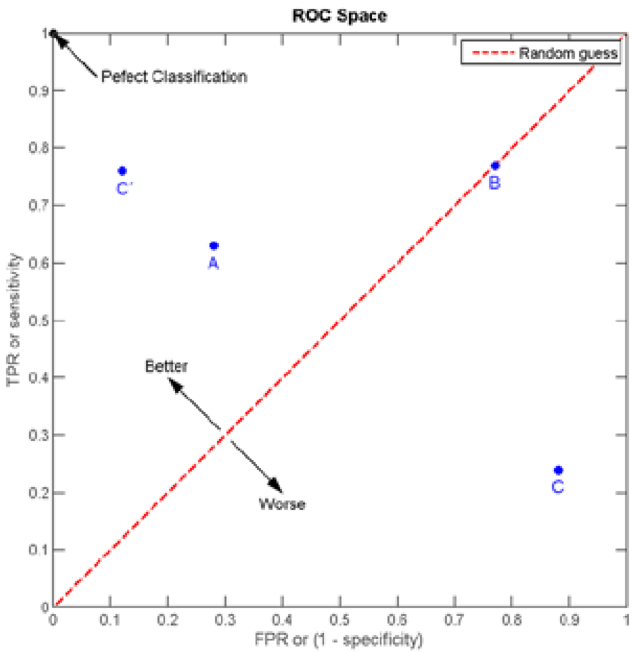


Figure 3. A sample ROC plot

The best possible prediction method would yield a point in the upper left corner of the ROC space. A random guess would give a point along a diagonal line from the left bottom to the top right corners. Points above the diagonal represent better than random classification results, while points below the line represent worse than random results. A sample ROC plot is shown in Figure 2. In general, ROC analysis is one tool to select possibly optimal models and to discard suboptimal ones independently from the class distribution. Sometimes, it might be hard to identify which algorithm performs better by directly looking at ROC curves. Area Under Curve (AUC) overcomes this drawback by finding the area under the ROC curve, making it easier to find the optimal model.

3. Results

3.1 Confusion matrix and ROC curve

Figure 4 shows the confusion matrix of the logistic regression model. The upper left region is true negative, the upper right region is false positive, the lower left region is false negative, and the lower right region is true positive. As shown in Figure 4, the logistic regression model has a relatively high (~67.0%) true positive rate and a relatively low (~32.4%) false positive rate.

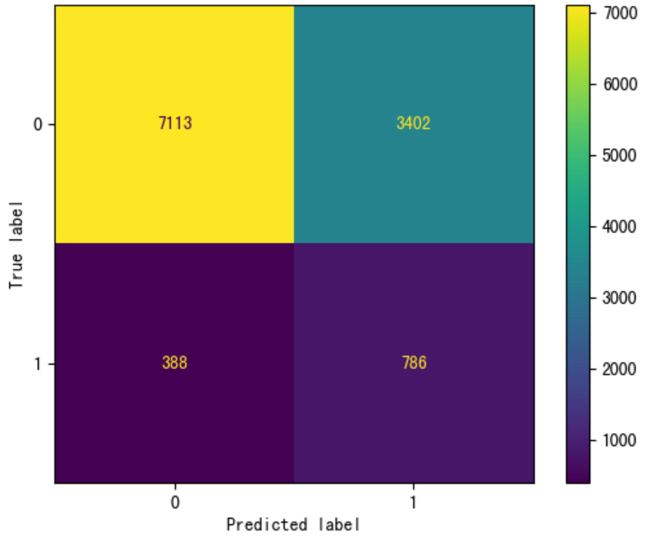


Figure 4. Confusion matrix of the predicted results

Figure 5 displays the ROC curve for the logistic regression model. It can be concluded that the model has results much better than random guessing and the AUROC score is 0.73.

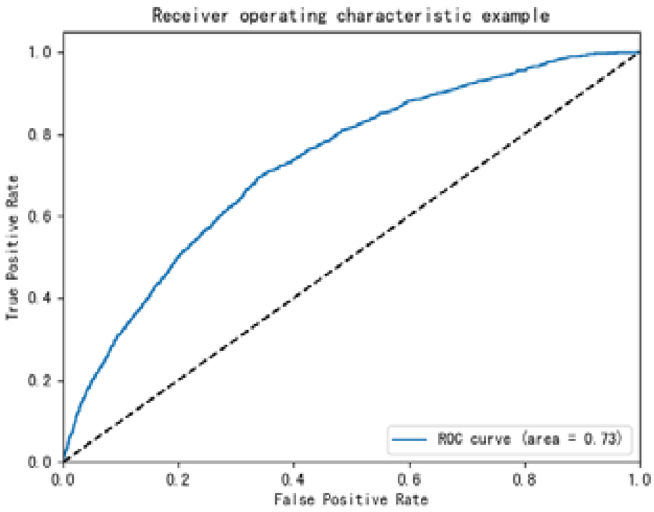


Figure 5. The ROC curve for the logistic regression model

3.2 Feature Importance

Like in linear regression, the coefficients in the logistic regression model also provide valuable information about the direction and magnitude of the impact of each input variable on the dependent variable. In other words, these coefficients can provide the basis for a crude feature importance score. The figure below shows the coefficient of each input variable.

The chart below shows A1_MENTHEALTH (parents' mental health) and ACE1 (hard to cover basic living expenses) are positively related to the development of ADHD, and ACE3 (parents not divorced) are negatively related to the development of ADHD. These results align with our findings from the correlation analysis. In addition, we also found that male children are more likely to develop ADHD and female children are less likely to develop ADHD (SC_SEX). This finding is coresonds with exisitng evidence suggesting that the prevalence of ADHD is greater in males than females [5] and ADHD is more commonly diagnosed in adult males compared with adult females at a ratio of 1.6:1 [6].

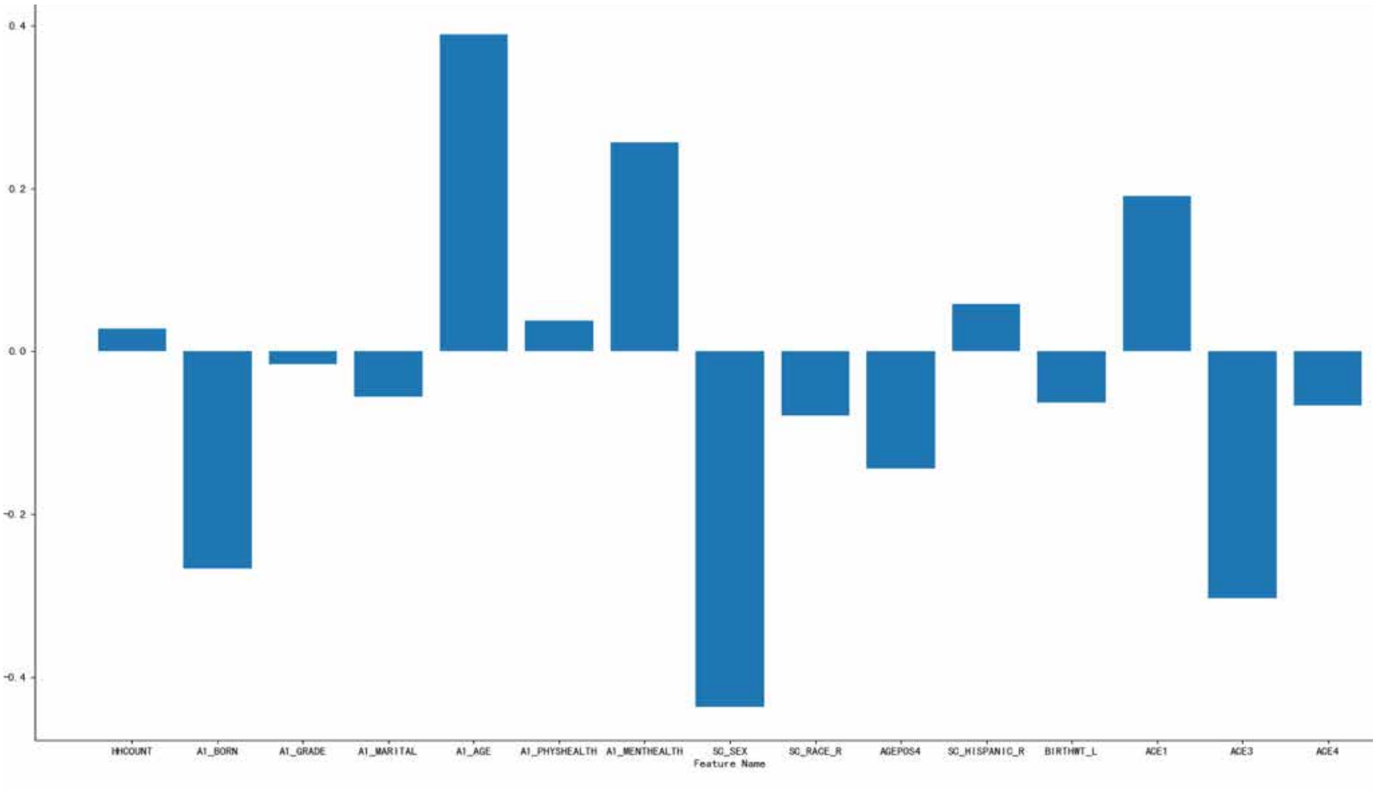


Figure 6. The importance score for each feature

4. Discussion

This study intends to build a predictive model to investigate the factors most related to the development of Attention Deficit/Hyperactivity Disorder (ADHD) among children. Through preliminary analysis, we discovered that parents' physical and mental health, family's financial ability to cover basic living expenses, as well as whether the parents are divorced are all risk factors. A logistic regression model was built, and the AUROC score is 0.73, indicating that the model has achieved relatively good performance in making accurate predictions on whether a child will develop ADHD. The predictive model suggests that A1_MENTHEALTH (parents' mental health) and ACE1 (hard to cover basic living expenses) are top risk factors for ADHD. A possible explanation of the results might be that a parents with worse mental health may have higher violence intention and may even abuse the child. In addition, children grown up in a family that is hard to pay for basic living expenses may be mentally insecure and thus are more prone to mental illnesses such as ADHD. This predictive model is helpful for healthcare professionals to identify children that are at higher risk for ADHD and come up with specific plans to reduce their risk for long-term impacts.

One limitation of this study is that we did not explore how individual independent variables have contributed to the overall predictive performance. Even though we can ascertain how each variable is correlated to our dependent variable through the model, we still have no idea how it influences the final model outcome. Therefore, this direction can be investigated in future studies. In addition, we only employed the vanilla logistic regression classification model in this study. Future studies can apply more complicated machine learning models, such as the artificial neural network, and compare its performance with logistic regression. With a highly accurate classification model, healthcare professionals might provide more customized and better service to children who are likely to have ADHD and find measures to minimize the long-term impacts on their development.

References:

1. NSCH 2003–2011: National Survey of Children's Health, telephone survey data; estimate includes children 4–17 years of age.
2. Danielson M. L., Bitsko R. H., Ghandour R. M., Holbrook J. R., Kogan M. D., Blumberg S. J. Prevalence of parent-reported ADHD diagnosis and associated treatment among U.S. children and adolescents, 2016. Journal of Clinical Child and Adolescent Psychology.– 47:2. 2018– P. 199–212.
3. Banerjee T. D., Middleton F., Faraone S. V. Environmental risk factors for attention-deficit hyperactivity disorder. Acta Paediatr.– Sep; 96(9). 2007.– P. 1269–74. Doi: 10.1111/j.1651-2227.2007.00430.x. PMID: 17718779.
4. Google. Classification: ROC Curve and AUC | Machine Learning Crash Course. Accessed November 25, 2021. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
5. Nøvik T. S., Hervas A., Ralston S. J., et al. Influence of gender on attention-deficit/hyperactivity disorder in Europe–ADORE. Eur Child Adolesc Psychiatry – 15(Suppl 1). 2006.– I15-I24.
6. Willcutt E. G. The prevalence of DSM–IV attention-deficit/hyperactivity disorder: a meta-analytic review. Neurotherapeutics – 9. 2012.– P. 490–499.