SAJEM EMERGENCY MEDICINE JOURNAL

ORIGINAL ARTICLE

Application of machine learning in predicting Attention Deficit Hyperactive Disorder (ADHD) in school going children of Pakistan

Salman Mansoor[1], Shoab Saadat[2], Sarah Noaman[3], Hamza Hassan Khan[4], Salman Assad[5]

Authors Affiliation:

University of Calgary, Alberta, Canada,[1]

Department of Nephrology, Mid-Essex Hospitals, NHS Trust, United Kingdom,[2]

Khyber Girls Medical College, Khyber Medical University, Peshawar, Pakistan,[3]

MBBS Graduate, Shifa College of Medicine, Islamabad, Pakistan,[4-5]

Correspondence to:

Salman Mansoor
salmanmansoor.dr@gmail.com

ABSTRACT

AIMS

The purpose of this study was to use machine learning algorithms to predict the probability of a child to have a certain attention deficit hyperactive disorder (ADHD) score under a given set of conditions.

METHODS

This was a cross-sectional survey which employed non-probability convenient sampling technique conducted at two schools in Islamabad, Pakistan. Using the latest version of Konstanz Information Miner (KNIME) Analytics, several machine learning algorithms were tested.

RESULTS

The area under the curve (AUC) for classification tree was 60.8% with a precision of 75.6% for the prediction of an ADHD score of 20 or more and the probability of 21.3% for a child to have an ADHD score of 20 or more. Important variables associated with a higher ADHD score included father's profession, school of the child, and the class of child.

CONCLUSION

This study shows that machine learning approach may be useful in developing a robust predictive model. Use of predictive model may allow use of limited resources towards assessment of children with higher probability of ADHD.

KEYWORDS

Attention deficit hyperactivity disorder (ADHD), Pakistan, Behavior rating scales, Machine learning approach

INTRODUCTION

Attention deficit hyperactivity disorder (ADHD) is characterized by pervasive and impairing symptoms of inattention, hyperactivity, and impulsivity.[1] ADHD is the most common psychiatric disorder of childhood (3% to 5% of children) with continued morbidity into adolescence and adulthood (50% to 70% of adults). Studies have shown that ADHD has a strong familial tendency. ADHD can cause a significant deficit in academic, social, and emotional functioning. Problems at school are a key feature of ADHD, often bringing the child with ADHD to clinical attention. It is important to establish the nature, severity, and persistence of school difficulties in children with ADHD.[2] Although ADHD in childhood was recognized in Diagnostic and Statistical Manual of Mental Disorders II (DSM-II) as hyperkinetic reaction of childhood almost 50 years ago, DSM-V only now has described the signs and symptoms of ADHD in adulthood.[3]

There is no definitive diagnostic test for ADHD, thus a comprehensive analysis should be made in multiple settings such as home, school, or at work by a healthcare professional. A wide array of rating scales, tests, and measures have been developed to aid in a systematic standardized assessment of various deficits associated with ADHD.[4]

The application of machine learning in predicting ADHD in school will help physicians in future to channelize appropriate treatment. In this study we will apply machine learning to predict ADHD incidence.

METHODOLOGY

Study Design

We employed a cross-sectional survey with non-probability convenient sampling technique conducted at two schools in Islamabad, Pakistan. The study included 500 children from grades 1-5. The questionnaire, in the Urdu language, was completed by parents based on their assessment of their child's behavior.

Interpretation of ADHD Questionnaire

The ADHD test and attention deficit disorder (ADD) test is based solely on behavior observations.[2] More than 20 checked items on this ADHD and ADD test indicate a strong tendency towards ADD or ADHD.

DATA ANALYSIS

Machine learning algorithms were used to generate a prediction chain for predicting the probability of a child to have ADHD score of 20 or more. Using the Konstanz Information Miner (KNIME) Analytics Version 3.5 and R version 3.5.1, several machine learning algorithms including classification tree, random forest, naive Bayes, and support vector machines were tested.

Explanatory variables like fathers profession in terms of either desk job and clerical job or skilled/general non-skilled labor/businessman/self-employed [Figure 1], school, class of the child and parent's interpersonal relationship were used as features for the final models. We included all the variables with at least 50% data availability which was the case for all explanatory variables. The outcome variable was whether a child scores more than 20 on an ADHD test or not. After complete data collection, data were divided into two subsets; training dataset (80 % of data) and the test dataset (20% of data). Next, using the training dataset, several machine learning prediction models were generated. The classification algorithms used included classification tree, random forest, naive Bayes, and support vector machines. This list of final candidate models was selected after a trial run on a minority data in the beginning which also included logistic regression. It was not found to be least accurate so was dropped out from the final list of candidate models. Once the models were generated, they were tested for accuracy on the remaining 20% of the dataset (test dataset) using 10 fold cross-validation.The model with predictions generating the highest area under curve was selected as better performing than the others.

RESULTS

Of the 500 children included in the survey, 290 were boys, 87 (17.4%) were in grade three, 190 (38.0%) were in grade four and 223 (44.6%) were in grade five. A majority (494/500) was between 8-10 years of age. There were 120 children who scored 20-29 while 17 had a score higher than 30. Thus, 24% students showed a strong tendency towards ADHD. More boys 81 (27.9%) scored 20 or greater than girls 39 (18.5%). Of the 500 children, 286 had fathers working as professionals and 385 children had stay-home mothers.

There was a significant relationship between the father's occupation and the child's score (p=0.032). There were 286 children in whom the fathers. Among them 232 (81.1%) had a score of (0-19), 48 (16.7%) scored (20-29) and 6 (2%) had a score of 30 and higher. There were 10 fathers who were skilled laborers, 5 (50%) among them scored (0-19) and 5 (50%) scored (20-29). None among the group skilled laborers scored 30 or above. Children in whom the paternal occupation was general labor there were 5 in total. Among this group 3 (60%) scored (0-19), 2 (40%) scored (20-29) and none scored 30 or above. There were 72 children in families where paternal occupation was a business. Among them 49 (68.1%) had a score of (0-19), 19 (26.3%) had a score of (20-29), and 4 (5.5%) scored 30 or above. In the group where the father's occupation was clerical, there were a total of 12 children. Among them 10 (83.3%) had a score of (0-19) and 2 (16.6%) scored (20-29) and none scored 30 or above. There were a total of 78 students whose fathers were self-employed. In these children 53 (67.9%) scored (0-19), 20 (25.6%) scored (20-29) and 5 students (6.4%) scored 30 or above [Figure 1]. About 385/500 parents had commented good inter-parental relationship ADHD was suspected in 91 (23.6%) while 115/500 had bad inter-parental relationship ADHD was suspected in 29 students (25.2%) [Table 1].

Characteristics		Scores		
		<19	20-29	>=30
Father's Occupation	Professional	232	48	6
	Skilled Labor	5	5	0
	General Labor	3	2	0
	Business	49	19	4
	Clerical	10	2	0
	Self Employed	53	20	5
	Other	28	7	2
	Subtotal	380	103	17

Mother's Occupation	House wife	298	76	11
	Professional	68	22	5
	Skilled Labor	4	2	0
	General Labor	2	2	1
	Clerical	3	0	0
	Business	1	1	0
	Self Employed	3	0	0
	Other	1	0	0
	Subtotal	380	103	17
Parent's inter-relation	Good	294	77	14
	Bad	0	0	0
	Separated	0	0	0
	No comments	86	26	3
	Subtotal	380	103	17

Table 1: Scores of the children calculated on the ADHD Test against parental characteristics

To explain further, the top performing model for predicting whether a child has an ADHD score greater than 20 was the one that used classification tree algorithm [Table 2].

Method Used	Training Data Set			Testing Data Set		
	AUC	Precision	Recall	AUC	Precision	Recall
Classification Tree	0.63	0.81	0.82	0.60	0.75	0.76
Naive Bayes	0.54	0.75	0.79	0.53	0.70	0.75
Random Forest Classification	0.62	0.83	0.83	0.59	0.80	0.80
Support Vector Machine (SVM)	0.55	0.85	0.81	0.48	0.82	0.80

Table 2: Algorithms used for predicting a child's ADHD score to be more than 20

Using this, we were able to predict with a precision of 75% and it gave an area under curve (AUC) of 60.8 %. This is reflected in the confusion matrix table [Table 3].

		Predicted ADHD Score		
		0-19	20-29	>=30
Actual ADHD Score	0-19	38	2	0
	20-29	5	2	15
	>=30	3	0	0

Table 3: Confusion Matrix of the test population (from the classification tree model) showing cross tabulation between the actual and predicted values of prediction of children having an ADHD score of 20 or above

This means that 75 out of 100 times, this model has the ability to correctly identify children with high ADHD scores. A classification tree was graphically generated (Figure 1) which demonstrated an overall probability of 21.3% for a child to have an ADHD score of 20 or more.

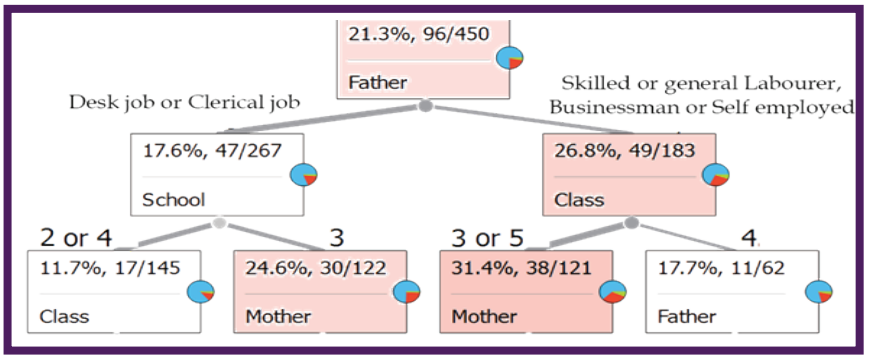

Figure 1: Classification tree model showing the most important predictors for an ADHD score of 20 or more

The next most important variables came out to be father's profession, school of the child and the class of child. The highest probability (31.4%) of a higher ADHD score was seen in children of class 3 or 5 who had working parents. Indeed, with more research and understanding of more explanatory variables, this precision power can be further enhanced using the modern machine learning techniques.

DISCUSSION

We found that the classification tree algorithm was the most accurate with an area under curve (AUC) of 60.8% and a precision of 75.6% for the prediction of an ADHD score of 20 or more.

To the best of our knowledge there haven't been previous studies which employed machine learning algorithms to generate a predictive model for children with ADHD in our region. This study will lead to other future surveys in employing new machine learning approaches for a more robust predictive model taking into account more variables in school going children having ADHD.

LIMITATIONS

The study failed to address further psychological testing and psychiatric treatment for the screened children due to limited resources. Self- reporting bias is another factor to be taken in account while generalizing the results. Localization bias where the study was confined to single institution should also be taken into account. The mother's survey responses were not addressed as they spend time with their children can also skew the dataset results.

REFERENCES

1. Polanczyk G, de Lima MS, Horta BL, Biederman J, Rohde LA. The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *Am J Psychiatry*. 2007; 164(6):942-8.

2. Loe IM, Feldman HM. Academic and educational outcomes of children with ADHD. *J Pediatr Psychol*. 2007; 32(6):643-54.

3. Kataoka SH. ADHD among U.S. children and adults: Increasing Access to Care. *Psychiatric Services*. 2016; 67(9): 937-937.

4. Hinshaw SP. Moderators and mediators of treatment outcome for youth with ADHD: understanding for whom and how interventions work. *Ambul Pediatr*. 2007; 7(1 Suppl):91-100.

5. Brown RT, Perrin JM. Measuring outcomes in attention-deficit/hyperactivity disorder. *J Pediatr Psychol*. 2007; 32(6): 627-630.