Predicting ADHD from early childhood data using data mining

Mohsen Nurisa (mohsen.nurisa@gmail.com)
Shiraz University https://orcid.org/0000-0002-3449-8359
Gholamhossein Dastghaibyfard
Shiraz University
Habib Hadianfard
Shiraz University

Research Article

Keywords: Artificial Neural Network, Data mining, Prediction of ADHD, Millennium Cohort Study, Strengths and Difficulties Questionnaire

Posted Date: April 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1395357/v1

Predicting ADHD from early childhood data using data mining

Mohsen Nurisa, Gholamhossein Dastghaibyfard, Habib Hadianfard

Department of Computer and Electrical Engineering, Shiraz University

Mohsen.nuirsa@gmial.com, 0000-0002-3449-8359

Abstract

One of the most common disorders among school-age children is attention deficit hyperactivity disorder (ADHD). Although the symptoms mostly appear between the ages of five and seven, some can be traced back to early childhood.

Objective: In this study, we identified and ranked early childhood characteristics that can correctly predict ADHD diagnosis at the age of seven based on none-clinical data, and then used those features in a computer model to enhance the prediction accuracy.

Method: The data used in this study was from the Millennium Cohort Study, which contains comprehensive information about the biological, genetic, and environmental characteristics of children and their parents. In our analysis, we conducted a complete mining process, including feature selection (regression and Support Vector Machine) and modeling (Artificial Neural Network) to select and use proper characteristics to predict ADHD diagnosis, and finally, evaluation (10-fold cross-validation) to assess the accuracy of the prediction.

Results: The proposed mining process selected and categorized 28 features (out of 3908) as the most important predictors, some[may] have not been reported by other studies before. These features belong to different age groups and both children and their parents. Total difficulty score of SDQ, child's weight, total health, and parent's income were among the features with the most predictive power. The results from the final model show an F1-score of 82.85%, which, compared to previous studies, shows a significant improvement.

Keywords

Artificial Neural Network, Data mining, Prediction of ADHD, Millennium Cohort Study, Strengths and Difficulties Questionnaire

Statements & Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Mohsen Nurisa. The first draft of the manuscript was written by [full name] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Introduction

Predicting mental disorders at early stages is crucial for planning interventions. This is especially true if it has a high prevalence. ADHD is one of the most common behavioral disorders among children and adolescents, affecting almost 6% of those under age 18 [1]. The serious impact ADHD has on the patient's life, as well as the financial burden it imposes on society, highlights the importance of early diagnosis and prevention [2].

ADHD is a developmental-behavioral disorder, symptoms of which fall into categories of attention deficit, hyperactivity, impulsive behavior, or a combination of them [3]. While symptoms usually disappear after puberty, in some cases, they persist into adulthood [4].

Data mining is a widely used technique to extract the relationship between events. Using this technique to extract the risk factors of ADHD and then predicting its occurrence can be an efficient and low-cost approach.

This study adds to the body of knowledge about features from early childhood that are meaningful in predicting ADHD. A comprehensive data mining process utilizes this information in a subsequent step to help achieve greater prediction accuracy than previous works.

This research includes five sections. The second section contains the research background, concepts, and vocabulary. The third section describes the current work in more detail and presents the data, the mining process, and the techniques used in this work. Finally, the results and the knowledge gained through this study are reported in sections four and five.

Literature Review

The factors predicting ADHD and identifying risk factors have been studied extensively. It has been shown that maternal conditions like hypertension during pregnancy (HDP) have links to ADHD diagnoses. [5]. Temperament traits in early childhood correlate with some ADHD symptoms in 10 years old children. There is a significant correlation between effortful control (EC) and anger with ADHD growth [6].

Many questionnaires have been used to predict ADHD. SDQ (Strength and Difficulties Questionnaire) is one of the questionnaires used for diagnosing and predicting ADHD. The sub-scales and the total score of SDQ filled out by parents when children were aged 5-7 years old have been used to predict whether the child would have ADHD at age 12. According to the results, SDQ filled in early childhood was a strong predictor of ADHD [7].

There has been extensive research on the predictors and risk factors of ADHD, but the time intervals considered in these studies were mostly limited to a specific period of child development. This study considered all stages of development up until the diagnosis of ADHD.

Based on the objectives and data in this study, a comprehensive data-mining process has been employed (prediction). The process includes stages such as data cleaning, source integration, feature selection, data transformation and preprocessing, modeling, and evaluation. Each step is discussed in detail in this section.

Methods

The database used is from the Millennium Cohort Study, which consists of eight different datasets related to different sweeps. Overall, it contains over 18,000 samples and over 9,700 variables. Table 1 briefly describes each of these datasets. We have used data from sweeps 1 through 3 as predictors (independent variables) and sweep four as the label (dependent variable).

Table 1 – Summary of datasets

No.	Dataset	No. of variables	Study – Child age period
1	UKDA-4683 (PI)	1732	1st – 1 to 9 months
2	UKDA-4683 (DV)	152	1st – 1 to 9 months
3	UKDA-5350 (PI)	3179	2nd study – 2 to 3 years
4	UKDA-5350 (DV)	205	2nd study – 2 to 3 years
5	UKDA-5795 (PI)	4127	3rd study – 3 to 5 years
6	UKDA-5795 (DV)	156	3rd study – 3 to 5 years
7	UKDA-6411 (PI)	not used	4th study – 4 to 7 years
8	UKDA-6411 (DV)	166	4th study – 4 to 7 years
	Total	**9717 variables**	

Generally, two types of information have been collected in this database: (1) the interviews conducted with the families, teachers, and children, and (2) assessments and measurements related to children. This database contains information about various topics including ethnicity, parent's income, housing, education, jobs, religions, and birth.

In this research, variables were categorized into three roles: identifier (1 variable), label (3 variables), and predictors. Since different algorithms require different types of data, in this stage, two numeric variables and one nominal variable were used as the label for ADHD:

1. ADHD: This variable is designated as DMADHDA0 in the fourth sweep (UKDA-6411-PI), which defines if a psychiatrist labeled the child with ADHD until the age of 7.
2. ADHDN: The variable 'DMADHDA0' converted into numeric (1 for Positive cases and 0 for negative cases)
3. HYPE: The variable 'DDHYPEA0' in the data from the fourth sweep, which denotes the score of children in Hyperactivity/Inattention subscale of QOL questionnaires. This variable is strongly correlated with the first variable. The results for the t-Test between this variable and the ADHD label is shown in Table 2.

Table 2 – Result of t-test for HYPE

		F	Sig.	t	df	Sig.(2-tailed)
HYPE	**With Equality of variance**	6/736	0/009	-26/000	13388	0/000
	Without Equality of variance			-27/692	182/470	0/000

Solution Overview

This section summarizes the six important stages of the mining approach and introduces the techniques used in each stage. Figure 1 illustrates the these stages.

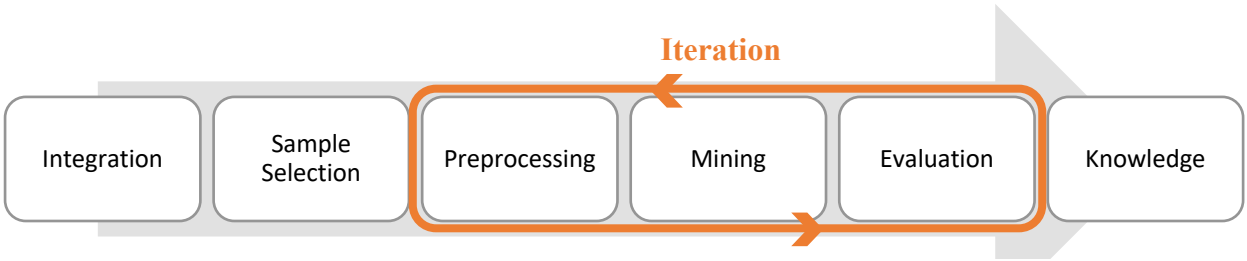

Fig. 1 Illustration of the mining approach

Source Integration: The database in use is comprised of eight different datasets. Each dataset contains a portion of the participants' data belonging to a specific period called a sweep. We used the uniquely identifier 'MSCID' to identify the samples (participants), then we merged their variables from across the database.

Sample Selection: Accordingly, only samples containing the 'ADHD' label were used in the analysis. This label holds the psychiatrist's opinion on the hyperactivity condition of each child and is from the 4th sweep. It should be noted that this label is available only for 76% (13,748) of the samples, and only 180 (0.1%) samples among them are 'positive'.

Preprocessing: Various transformations were required before raw data was ready for modeling. There were two reasons for these transformations. First, the database had too many variables. Secondly, the label variable was imbalanced (the high portion of negative samples against few positive ones). Furthermore, since these data have been collected through seven years, the trend and changes in some variables might've contained hidden knowledge. Therefore, various methods of feature selection, sampling, and feature generation was used.

Modelling: After the preprocessing, 286 samples and 28 features have been selected to create the models. Regression and Artificial Neural Networks are two of the used models and algorithms in this stage.

Evaluation: Various metrics such as precision, recall, and mean square error were used to evaluate the performance of each method.

Knowledge Representation: The weighting of the features was done according to their importance in recognizing and discovering effective factors for ADHD in five-year-olds. Some of these variables can predict signs of ADHD as early as nine months after birth.

Feature Elimination: The preprocessing stage starts with integrating all of the datasets and joining them to create a single view. The resulting tables contained over 10,000 variables. The datasets usually contain a large volume of useless information because most of the variables are irrelevant. The process of removing unnecessary variables is called feature elimination and was done through two steps:

1. Remove all numeric variables with a deviation equal to 0 (all values being the same).
2. Remove nominal variables whit more than 95% identical values.

The elimination process removed over 69% (6,724) of variables in the dataset.

Feature Generation: The value of most of the variables changes over time. According to the descriptions and names of the variables, we identified similarities between 905 feature pairs. We calculated the difference for each feature pair to obtain the trend of changes. Table 3 shows the number of feature-pairs in each pair of sweeps.

Table 3 – Number of repeated variables in the dataset

Prior Study	Sub. Study	No. of pairs
1	3	242
2	3	394
1	2	269
Total		**905**

Table 4 shows the rank of each generated feature based on its importance (Information-Gain). Based on this score, BMI of the mother, BMI of the child, and other weight-related variables were the most important trends related to ADHD.

Table 4 – Info-Gain weights of generated features

No.	Prior Study	Sub. Study	Variable Pair	Description	Weight
1	1	2	BDMBMI00 ADMBMI00	Mother's BMI	0.72
2	2	3	CMDBMI00 BMDBMI00	Child's BMI	0.66
3	2	3	CDMBMI00 BDMBMI00	Mother's BMI	0.66
4	2	3	CMGROA00 BMGROA00	Parents Gross Income	0.55
5	1	2	BMWGTK00 AMWGTK00	Child's Weight (Kg)	0.51

Missing Values: Real-world data often comes with missing values. In some cases, the value for these variables can be estimated, and in other cases, they might be replaced with other values such as the average value. The datasets used for this study contained some missing values too. For example, the value of the variable for "Mother's illness before birth" can be '-9', meaning they refused to answer this question, or '-8' meaning they didn't know.

In this study, missing values were replaced with the average values, and for nominal variables, they were replaced with the most common value. Even though this is a simple workaround, it can still be used when there are too many variables or if complex models cannot be applied. As for this database, from the overall 3,908 features (2,998 original features and 910 generated ones), 207 of them contained at least one missing value.

Label Balancing: As mentioned, only less than 0.1% (180) of the cases had a positive ADHD label. Most of the prediction and feature selection algorithms cannot handle this amount of disproportion. Therefore, the cases had to be selected in a way that accurately represented both classes.

Different studies have chosen different ratios of positive and negative cases to balance the label stratification. In this study, 1:1 and 1:2 (Twice as many negative labels) ratios were compared, and it was found that 1:1 was the most efficient ratio.

Modeling

Feature Selection In data mining, feature selection can be an automatic or manual procedure of selecting the variables with the highest impact on the target variable. The presence of irrelevant variables might slow the model down or decrease its accuracy. In this study, different feature selection methods were applied, and a subset of the features from each method was chosen for the final feature set.

An overview of the steps involved in feature selection is shown in figure 2. The rest of this section covers some of these methods.

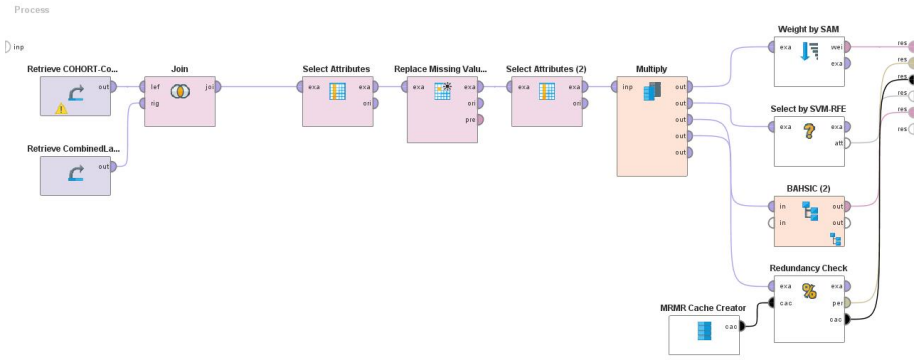

Fig. 2 Feature selection process and it's steps

Linear Regression Regression models are one of the most common methods for numerical prediction. In this work, for different subsets of variables, multiple regression models (Ensemble) were created.

The final mean square root error value for this model was 11.001. Figure 3 shows the weights for predictors. Based on these results, the variables from sweeps 2 and 3 (variables starting with letters B and C) have the most impact on the target variable. 'CDHYPEA0' (hyperactivity/inattention subscale of the SDQ at the age of 3-5) and 'CDEBDTA0' (the total difficulties subscale of the SDQ) were the two most important ones.

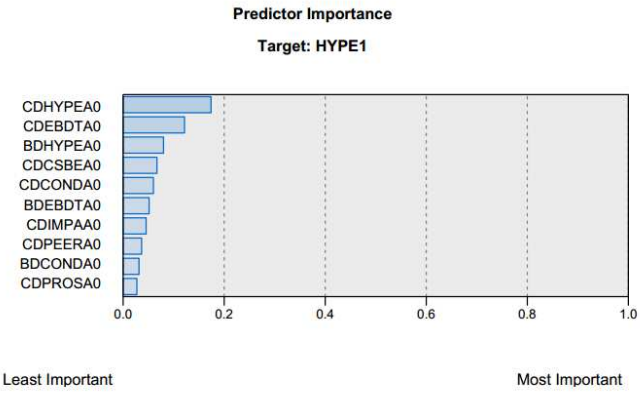

Fig. 3 The relative importance of each variable in Linear Regression model

Evolutionary Feature Selection

The evolutionary feature selection method is a method inspired by natural selection. It starts by selecting p variables; then, based on the evaluation score at each iteration, it adds or removes some variables as it creates the model. After each iteration, the best models get chosen to be the initial set for the next iteration.

In this work, a linear regression model with the p-value of 10was used. To validate the models, the ratio of test to train cases was 3:7. Because regression models can only handle numerical data, in this method only numeric variables were included. Table 5 shows the results of these tests.

Table 5 – 18 Selected variables by Evolutionary Feature Selection

Variable Name	Description	Variable Name	Description	Variable Name	Description
AMPDBM00	Date of Birth (month)	AMADMOA0	Number of health problems	AMPUWK00	Number of units in average week before pregnancy
AMDMDAA0	Age first had formula milk	AMACCAA0	Number of accidents or injuries	AMHARE00	Happy/Unhappy with relationship
AMCMMTA0	Age first had cow's milk	AMWEIS00	Current weight	AMNETA00	Take-home pay last time
AMSFMTA0	Age first had solid food	AMSMMA00	Number of cigarettes currently smoked per day	BPFFSP00	Partner has sight problems
BDCONDA0	SDQ Conduct Problems	BDHYPEA0	SDQ Hyperactivity/Inattention	COEDEX00	weekly net family income
CDHYPEA0	SDQ Hyperactivity/Inattention	CDEMOTA0	SDQ Emotional Problems	CDEBDTA0	SDQ Total Difficulties

Using an evolutionary feature selection method can drastically improve the performance of regression models. Fig. 4 illustrates the decreasing rate of model error with each generation (iteration). According to the results, the error of 11.001 has been reduced to 2.351.

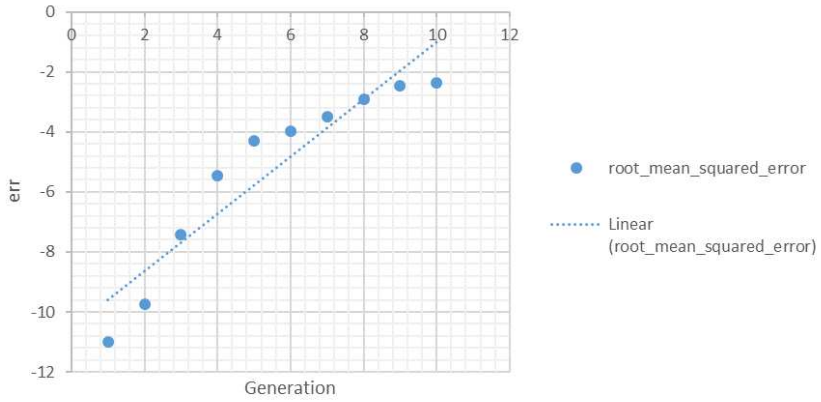

Fig. 4 *Changes of root mean squared error after each Generation*

Models We used Artificial Neural Network (ANN) algorithms as the final model. A neural network consists of an interconnected group of artificial neurons that process information using weighted links. In each iteration, the model changes the weights of each neuron to push its output closer to the target.

The rate at which ANNs modify the weights is called the learning rate. In this study, a value of 0.3 was used for this parameter. The model was trained using the 28 features chosen in the feature selection stage. The results are reported in detail in section 4.

Evaluation

We used 10-Fold-Cross-Validation to assess the validity of our result. This method works by dividing the population into 10 groups (folds). It uses nine groups as a training set and the remaining one as testing set in each iteration. The result is calculated by taking the average of all ten tests.

A comparison of the predicted values with the actual values allowed us to assess the accuracy of the classifier. Since both the predictions and original numbers each can have two different values (Positive and Negative), this results in a combination of four outcomes. A matrix called the 'Confusion Matrix' displays these combinations which are illustrated in Figure X.

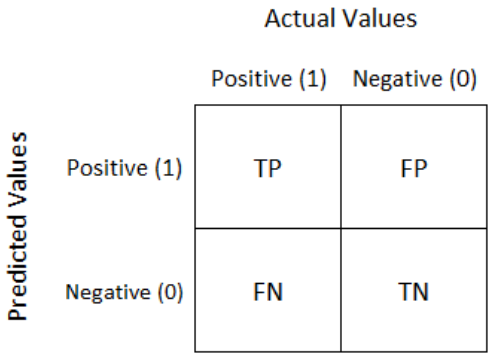

Figure 5 *Confusion matrix*

To evaluate the results based on the confusion matrix, two metrics were created, precision and recall. The precision of the test measures the proportion of true positive cases among all positively classified cases, and recall measures the percentage of true positive cases among all positive cases reported in the data.

The confusion matrix for the ANN model is presented in Table 6.

Table 6 – Confusion matrix for ANN model

	true No	true Yes	class precision
pred. No	122	27	**81.88%**
pred. Yes	21	116	**84.67%**
class recall	**85.31%**	**81.12%**	

Results and Discussion

Temporal Classification of the variables There are different ways to categorize the variables used in this study. One of these categories is by time, which means how much each sweep contributes to predicting ADHD (label) at age seven. Figure 5 shows the frequency distribution of predictor variables at any given period.

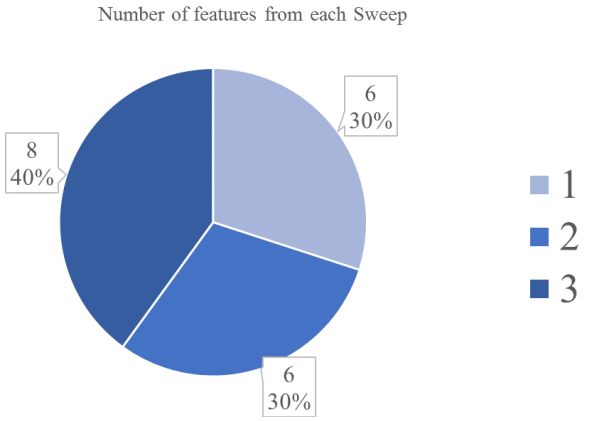

Fig. 6 *Number of important features in each sweep*

As mentioned before, for sweep 1, data is collected between two and eight months of age. Additionally, sweeps 2 and 3 reflect two to three-year-olds and three to five-year-olds, respectively. Figure 5 shows that among these three sweeps, the third one has the largest volume of related features. In other words, most of the features with higher prediction power belong to the three to five-year-age period.

Since this chart cannot assert anything about the predictive power of any of the variables, it is not possible to indicate which section is more involved in predicting the label. Therefore, although the features of sweep 3 constitute only 43% of the entire features, their prediction power may be more or less. However, it is still worth considering that the neural network model created is largely dependent on data collected in sweeps 1 and 2. These variables are often about early childhood.

Figure 6 shows the results related to the prediction power of the variables of each sweep based on the SAM algorithm. In this diagram, the prediction power and importance of each sweep are nearly equal to the number of variables they have in the model.

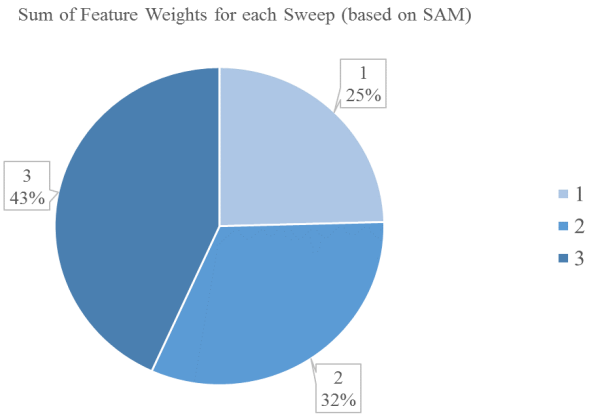

Fig. 7 *Relative importance of features in each sweep*

Conceptual Classification of the variables The variables selected in this study were related to various factors. At the macro level, these factors include SDQ questionnaires, measurements of children, or parent interviews. These variables were associated with age, income, health, and other factors at lower levels.

A cluster analysis was used to describe the characteristics of variables. The income and SDQ scores are the most significant categories, as depicted in Figure 7. 11 variables also were categorized as "Other".

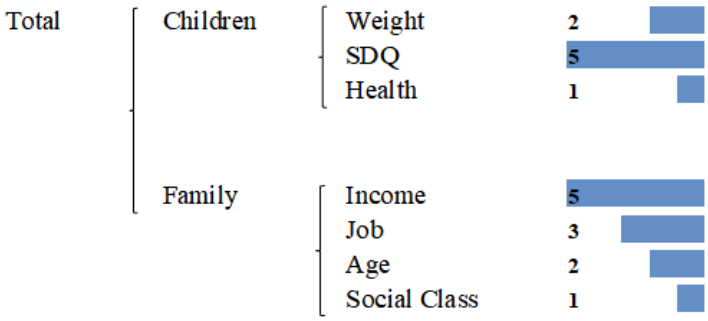

Fig. 8 Categorization and Distribution of selected features

Comparative Analysis Predicting ADHD in children has always been the goal of many previous studies. In this regard, similar to the present study, many studies have explored factors influencing the prediction of ADHD. Most of the previous studies have used the results of regression tests to assess the factors identified.

Since in this study, the discovery of factors was associated with the creation of a prediction model, to be able to compare the results, we evaluated the pre-processed data and selected variables through regression models. Table 7 displays the results of this test model with 26 variables.

Table 7 – Summary of Regression model using selected features

Model	R	R^2	Std. Error
	0.782	0.611	0.32
Predictors: (Constant), CMMOAD00AMMOAD00, AMSOCC00, CDIMPAA0, BMLIYR00AMLIYR00, CSCHIDAA, AMLWTGA0, CMMPNXAA, BMMOAD00AMMOAD00, BPJBSO00, BMFJSC00, AMNETA00, AMDAGB00, BDMPCFA0, CDHYPEA0, CDOEDE00, CDPEERA0, AMGROA00, BDHYPEA0, CDCONDA0, ADVMCEQINC, BDEBDTA0, COEDEX00, CDEBDTA0, BMPDBY00, AMDAGI00, BMDAGB00			

The results displayed in Table 7 indicate that the predictors can explain 61% of the variance observed in the study variable. Previous studies have used a wide range of variables to perform their analysis. However, the results of this study show improvements compared to previous studies. For example, the value of R2 compared to Einziger et Al has now been improved by 0.2 and the value of Specificity (Recall) has improved by 40% compared to Rimvall.

Some of the variables selected in this study are similar to those in previous works. For example, Foulon et al also reported that family income levels, smoking, and alcohol consumption of parents, breastfeeding during infancy, and family health conditions were also associated with ADHD. However, as shown, a number of variables in this study are new.

Conclusion

The objective of this study was to develop a method for predicting the diagnosis of ADHD based on early childhood data. The Millennium Longitudinal Cohort Study was used to resolve the lack of integration and data from various stages in a person's life. It consists of several sweeps and contains data about people of different age ranges, including their environment, genetics, and personal characteristics.

A large number of variables and repeated studies over time resulted in a large number of repetitions, insignificant and missing data in this data set. So, we had to clean the data before we could begin analyzing it. The data mining process began with 3908 variables, and after several stages of data purification and pre-processing, 28 variables were finally selected for the final model. Using these variables in an artificial neural network yielded a good prediction power. We also categorized these variables into several categories, which included age groups. We found that even though most of the variables belonged to the age range of 3-5, almost 25 percent belonged to early childhood where children were 2 to 8 months old.

This study showed that ADHD symptoms can be accurately predicted years before they appear. The results can be used to guide the collection of children's information.

References

[1] Polanczyk, G., De Lima, M. S., Horta, B. L., Biederman, J., Rohde, L. A.:The worldwide prevalence of ADHD: A systematic review and metaregression analysis. *Am. J. Psychiatry* (2007). https://doi.org/10.1176/ajp.2007.164.6.942.

[2] Chorozoglou, M., Smith, E., Koerting, J., Thompson, M. J., Sayal, K., Sonuga-Barke, E. J. S.:Preschool hyperactivity is associated with long-term economic burden: Evidence from a longitudinal health economic analysis of costs incurred across childhood, adolescence and young adulthood. *J. Child Psychol. Psychiatry Allied Discip (*2015*)*. https://doi.org/10.1111/jcpp.12437.

[3] Wilens, T. E., Spencer, T. J.:Understanding Attention-Deficit/Hyperactivity Disorder From Childhood to Adulthood. *Postgrad. Med.* (2010*)*. https://doi.org/10.3810/PGM.2010.09.2206.

[4] Rucklidge, J. J.:Gender Differences in Attention-Deficit/Hyperactivity Disorder. *Psychiatr. Clin. North Am* (2010). https://doi.org/10.1016/j.psc.2010.01.006.

[5] Böhm, S. , Curran, E. A., Kenny, L. C., O'Keeffe, G. W., Murray, D., Khashan, A. S. :The Effect of Hypertensive Disorders of Pregnancy on the Risk of ADHD in the Offspring. *J. Atten. Disord* (2019). https://doi.org/10.1177/1087054717690230.

[6] Einziger, T. , *Levis, L., Zilberman-Hayun, Y., Auerback, J.G., Atzaba-Poria, N., Arbelle, S., Berger, A.*:Predicting ADHD Symptoms in Adolescence from Early Childhood Temperament Traits. *J. Abnorm. Child Psychol* (2018). https://doi.org/10.1007/s10802-017-0287-4.

[7] Rimvall, M. K., Elberling, H., Rask, C. U., Helenius, D., Skovgaard, A. M., Jeppesen, P.:Predicting ADHD in school age when using the Strengths and Difficulties Questionnaire in preschool age: a longitudinal general population study. *Eur. Child Adolesc. Psychiatry* (2014). https://doi.org/10.1007/s00787-014-0546-7.