International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

DEVELOPING A PREDICTIVE MODEL FOR EMOTIONAL AND BEHAVIORAL DISORDER AMONG CHILDREN

Yicen Tao

Rutgers Preparatory School, New Jersey

DOI: 10.46609/IJSSER.2022.v07i04.009 URL: https://doi.org/10.46609/IJSSER.2022.v07i04.009

Received: 1 April 2022 / Accepted: 15 April 2022 / Published: 30 April 2022

ABSTRACT

Emotional and behavioral disorder (EBD) has been an ongoing issue in the United States. According to CDC, about 17% children and teenagers are reported to have symptoms of EBD. Thus, it is imperative for groups like parents, guardians, doctors, and teachers to develop a more in-depth understanding of the causes that lead to EBD. The data in this study are the responses of 67,625 individuals to the survey National Survey of Children's Health (NSCH).Mean value imputation, and min-max scaling are some of the pre-processing techniques that we used to prepare the data set for later model-building. Then we selected some independent variables including demographic information, and special therapy status, and used them to develop two predictive models, a logistic regression model and an artificial neural network, in order to predict the probability of developing EBD. Furthermore, we used AUC and ROC to evaluate the accuracy of the two models. Both predictive models have good performance. The AUC of the two predictive models are 0.93 and 0.94, respectively, and the ROC curve of the models are also good and similar. The results indicate that older children are more likely to have EBD, while the children who are receiving or have received special therapy for mental problems are less likely to develop EBD. This report concludes that offering quality health care to the children in need is an effective way to decrease their probability of developing EBD.

Keywords: emotional and behavioral disorder, logistic regression model, artificial neural network

1. Introduction

Emotional and behavioral disorder is a broad term including various types of disorders. The most common ones are ADHD, anxiety, antisocial behaviors. People with emotional and behavioral

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

disorder (EBD) have symptoms that varies from one another, and children's symptoms can be observed in academic performance as well as social ability. Emotional and behavioral disorder among children under the age of 18 is not a rare problem. According to CDC, among children aged 3 – 17, 7.4% have behavioral disorder, 7.1% have anxiety, and 3.2% have depression (CDC, 2021). Children with emotional and behavioral disorder often find themselves hard to have a good relationship with others or difficult to perform well in schools. They typically have either internalizing behaviors or externalizing behaviors. For children with internalizing behaviors, they tend to direct the negative mood, including anxiety and depression, toward themselves. Externalizing behaviors, on the other hand, are actions that direct negative energy toward surroundings. Children with externalizing behaviors often act violently, such as breaking the rules and doing activities with physical violence. Both behaviors are problematic, either to themselves or to the society. For example, in schools, children with emotional and behavioral disorder will significantly affect other students both academically and socially.

Much research has been done in the field of emotional and behavioral disorder to find out its possible causes. A group of researchers from Naikai University (Wang et. al., 2021) used adolescent mice to establish the relationship between EBD and the exposure to melamine cyanuric acid, and they concluded that the exposure would cause mice to have depressive-like and anxiety-like behaviors, which are typical symptoms of EBD. Another research by Dolores Garcia-Arocena listed some potential indicators to emotional and behavioral disorder such as Serotonin and Dopamine. Moreover, the research provided several therapies and how each one of it could treat the problem. Although the research by researchers from Naikai University and Dolores Garcia-Arocena contributes to our understanding between EBD and certain chemical compounds, it does not help us diagnose EBD among children in an early stage effectively.

A precondition of the medical treatment is to recognize the existence of emotional and behavioral disorder on children with possible symptoms. In many cases, when children behave inappropriately, parents will not immediately consider the possibility of EBD. To help parents better evaluate the possibility of EBD, in this study, we are going to examine the predictors of emotional and behavioral disorder medication and build a predictive model for it using the logistic regression model and the artificial neural network.

2. Method

2.1 Data

This report uses data from the National Survey of Children's Health (NSCH) in 2019, which is a population-based survey established by the Health Resources and Services Administration

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

(HRSA) Maternal and Child Health Bureau (MCHB)to monitor the prevalence of the children health condition in the United States and to evaluate their access to quality health care (NSCH - Questionnaires 2019). The whole survey mainly encapsulates family composition, children sex, special therapies, current medication, and a list of other health and family related questions. The data is collected from random households in the US by telephone surveys. The 2019 NSCH dataset is used in this report. Before the data-cleaning process, the NHIS dataset has 67,625 valid observations.

The table below shows all the variables that have been chosen in this report to examine the relationship between independent variables and the dependent variable:

Table 1: Variables used for analysis

Item Code	Question	Function
TOTKIDS_R	Number of Children in Household	Independent Variable
TENURE	The Conditions under Which Land or Buildings Are Held or Occupied	Independent Variable
MPC_YN	Metropolitan Principal City Status	Independent Variable
C_AGE_YEARS	Child Age	Independent Variable
C_RACE_R	Race of Child	Independent Variable
C_SEX	Child Sex	Independent Variable
C_K2Q10	Child Needs or Uses Medication Currently	Independent Variable
C_K2Q16	Child Limited Ability	Independent Variable
C_K2Q19	Child Special Therapy	Independent Variable
C_CSHCN	Special Health Care Needs Status of Child	Independent Variable
C_FWS	Child Weight	Independent Variable
C_K2Q22	Child Needs Treatment for Emotion Develop Behave	Dependent Variable

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

This report uses the variable "C_K2Q22" as the dependent variable. Responses to the question "C_K2Q22" is dichotomous, meaning that the respondents either answer "yes", indicating that the child needs treatment for EBD, or "no", indicating that the child does not need such treatment.

2.2 Statistical Models

2.2.1 Pre-processing

Some pre-processing techniques are used in this step to improve the accuracy of this data set. Since there is inevitably missing data, imputation is required to better analyze and extrapolate the missing data. Due to the defect of most machine learning algorithms that missing values could not be processed, we use the mean value imputation to fill in the missing values using the mean value of the column. As required by some machine learning algorithms, such as the one we in the report called artificial neural networks, we use feature scaling to convert different data into comparable scales to improve the accuracy.

In this report, we will use the min-max scalar for this purpose. For each feature, the maximum and minimum are computed as X_{max} and X_{min}. Then each data point X with respect to that feature is replaced by X_{sc} calculated as:

$$X_{sc}=\frac{X-X_{min}}{X_{max}-X_{min}}$$

Using this formula, X_{sc} is the ultimate value that is going to be analyzed in this report.

2.2.2 Logistic Regression Model

A logistic regression model refers to a model that is used to predict the probability of an incidence to happen. The probability varies from 0 to 1, with zero indicating not likely to happen and one indicating very likely to happen. Instead of a linear relationship, the logistic regression model fits an "S" shape which can be expressed using the formula below:

$$\ln\left(\frac{y}{y-1}\right) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

In the above equation, a_0 is the intercept,x_n represents the independent variables, and $a_1$to a_n, are their corresponding coefficients (weights). In this report, our goal is to find the coefficients (a_0...a_n) minimizing the sum of squared errors (SSE) so that our predicted values will deviate the least from the real values.

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

2.2.3 Artificial Neural Networks

An artificial neural network is a system that shows the interrelation of each variable (input) to the result (output) through layers and nodes. The system is inspired by the biological neural networks that the inputs will travel through the hidden layers when signaled and eventually send the information to the output layer. Unlike in biological systems, here the "signal" is a real number, and the output of the neurons could be computed while the sum of some non-linear functions has applied on the inputs.

In a typical artificial neural network, there are a input layer, several hidden layers, and a output layer. The input payer receives the data; the hidden layer process the data; the output layer transformed it into some predicted labels. In this report, the artificial neural network model consists one hidden layer with three nodes.

In each layer, there are also edges connecting the nodes from the previous layer to the nodes in the current layer, and those edges are often used as weights during the calculation process. Like in logistic regression models, the goal for training artificial neural networks is to find a set of edges (weights)that minimize our cost function and to achieve the best prediction performance.

A package called “neuralnet” in R was used to conduct neural network analysis (Fritsch et. al, 2019). The package neuralnet focuses on multi-layer perceptron, which is well applicable when modeling functional relationships.

2.3 Model Validation

The true positive rate (TPR) and false positive rate (FPR) need us to get four values before calculation. True positive (TP) is when the prediction outcome and actual value are both positive. True negative (TN) has both values negative. False positive (FP) is when the prediction outcome is positive but the actual value is negative, and false negative (FN) is when the prediction is negative but 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:

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

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

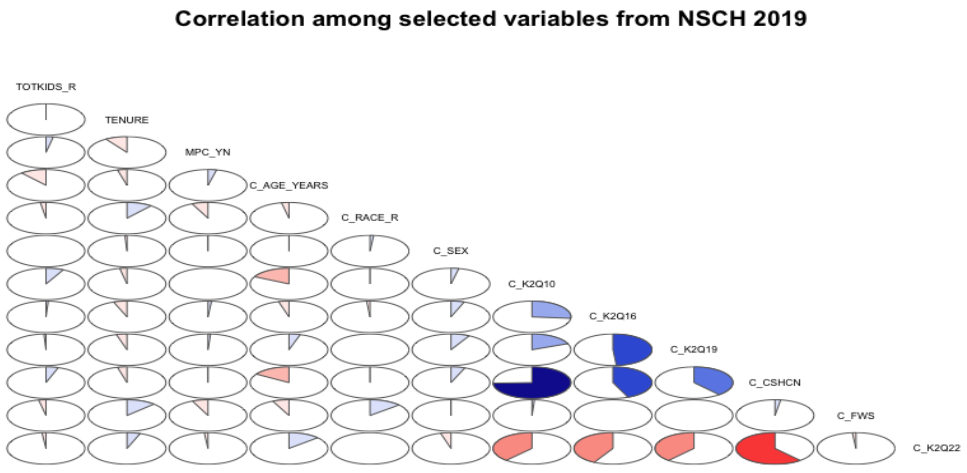
A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied (Google, 2020). In ROC curve, the best prediction method would have a point showed in the upper left corner. The diagonal line from the top right corner to the left bottom corner represents a random guess. So, any random guess would have a point on the diagonal line. If a point is above the line, it means that the method is better than to randomly classify. Conversely, a point below the line represent that the method is worse than random classification results. Overall, ROC curve analysis tests the models and allow us to select the optimal one. However, it is possible that we could not identify the optimal one only by looking at ROC curves. So, we use Area Under Curve (AUC) to help us select the better model.

3. Results

3.1 Chorogram

A chorogram can show the correlation of two variables by presenting cells. The cells can be filled in with different colors and shade to represent the signs and magnitudes of their correlations. In this report, blue represents positive correlation, and red represents negative correlation. And darker color represents closer correlation.

Figure 1: Correlation among variables



According to the chorogram above, children's need for EBD-related treatment has the strongest positive correlation with their age and has the strongest negative relationship with "C_CSHCN",

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

"C_K2Q10", "C_K2Q16", "C_K2Q19", which is a series of variables indicating their need for special care.

3.2 Logistic Regression Results

The results of logistic regression analysis of children ever need medical treatment for emotional and behavioral disorder are listed in Figure 2.

Figure 2: Logistic regression results

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.226776   0.106376  -2.132  0.03302 *
TOTKIDS_R    0.239056   0.076000   3.145  0.00166 **
TENURE       0.577085   0.086196   6.695 2.16e-11 ***
MPC_YN      -0.091501   0.062322  -1.468  0.14205
C_AGE_YEARS  1.606550   0.100896  15.923  < 2e-16 ***
C_RACE_R    -0.040904   0.085995  -0.476  0.63432
C_SEX        0.008856   0.049104   0.180  0.85688
C_K2Q10      0.693631   0.061743  11.234  < 2e-16 ***
C_K2Q16     -0.582991   0.066924  -8.711  < 2e-16 ***
C_K2Q19     -1.103371   0.065954 -16.729  < 2e-16 ***
C_CSHCN     -4.576340   0.084993 -53.844  < 2e-16 ***
C_FWS        0.260209   0.675181   0.385  0.69995
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the logistic results, it is not hard to find that, taking a 99.9% confidence level, family's occupation condition (i.e. whether the house is rented or owned), child's age, and the need for special care are all significant predictors of the dependent variable. More specifically, older children are more likely to develop EBD, while those who have received or are currently receiving special care are less likely to develop EBD.

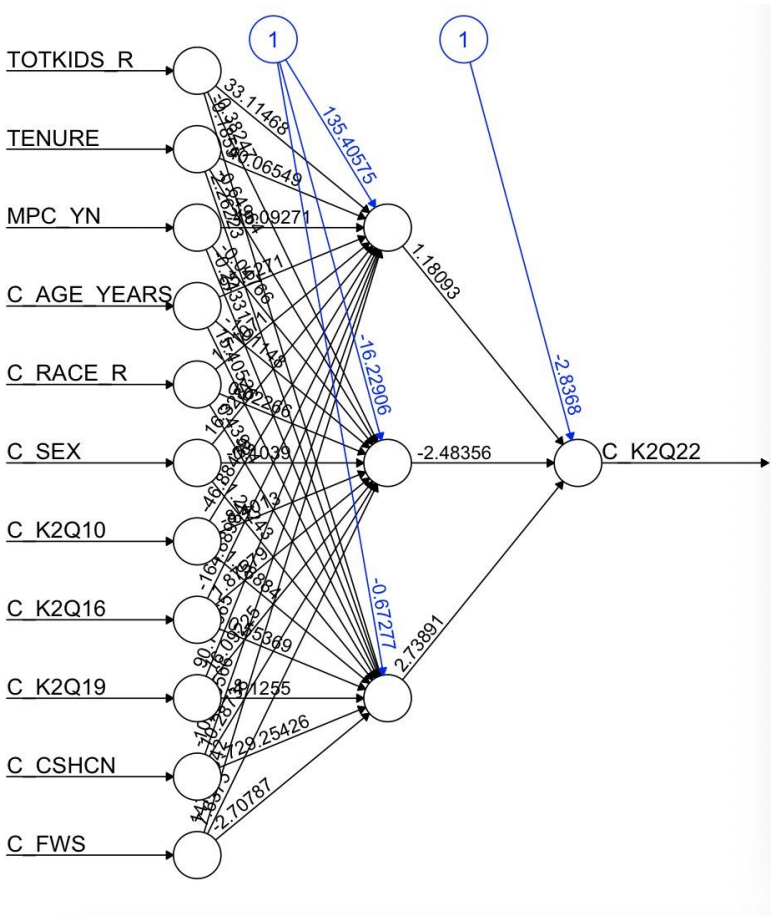
3.3 Artificial Neural Network Results

Figure 3 presents the structure of artificial neural network. The number around the arrow represents the corresponding weight.

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

Figure 3: Structure of the artificial neural network

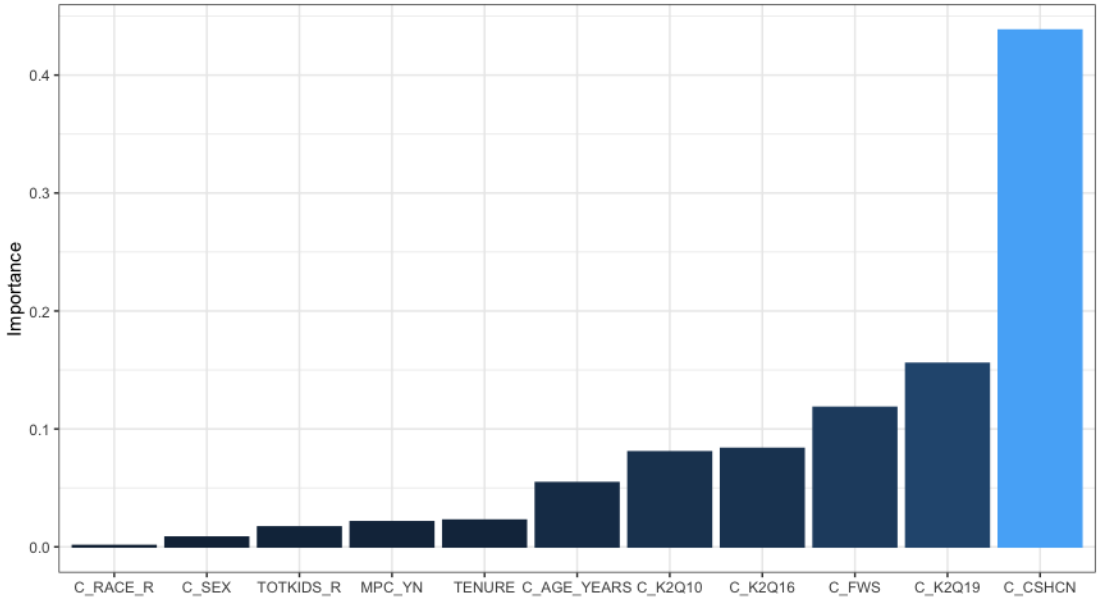


To figure out the relative importance of independent variables, Garson describes a method that can be used to identify the relative importance of a group of independent variables for a dependent variable in an artificial neural network (Garson, 1991).Whether the independent variable is important to the dependent variable or not, can be figure out by identifying the weighted connections of the nods. This process will be repeated for all independent variables until we get all weights of Figure 4 shows the importance of each question using Garson's algorithm.

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

Figure 4: The importance of each question in the artificial neural network.



The most important predictor is the respondent's special care need status, followed by the need for special therapy, weight, and whether the respondent is of limited ability.

3.4 Model Validation

Figure 5 displays the ROC curve for the logistic regression model and the artificial neutral network model, and Table 2shows the respective AUC scores of both models. Combining both Figure 5 and Table 2, it can be concluded that the performance of both models are rather similar, while the artificial neural network being slightly better than the logistic regression.

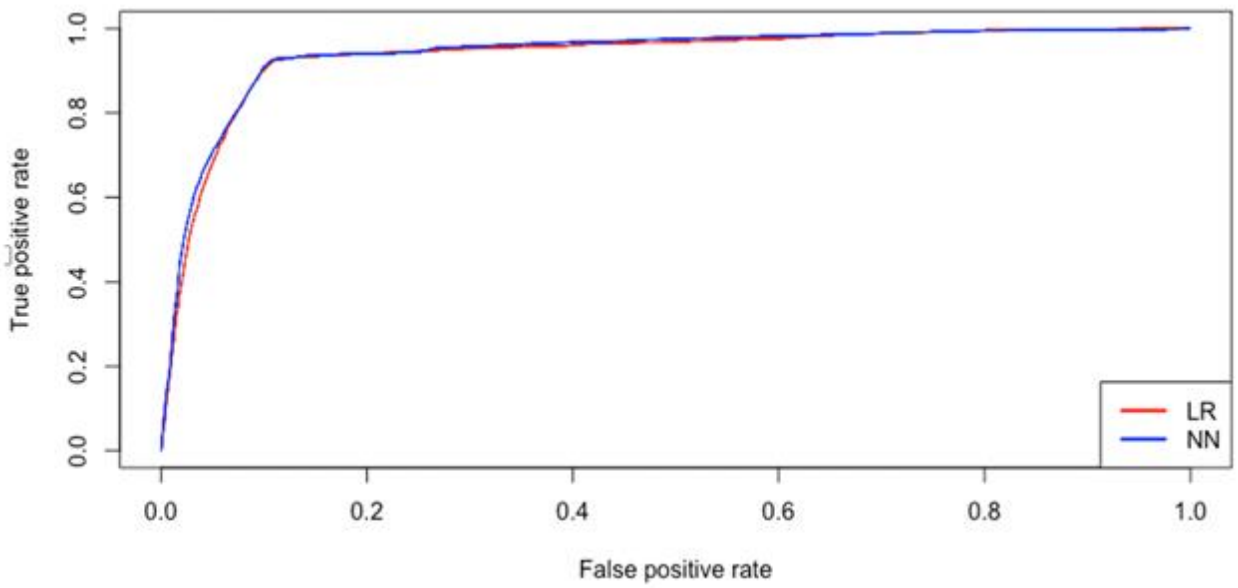
Table 2: The AUC score for logistic regression and artificial neural network

Algorithm	AUC Score
Logistic Regression	0.93
Artificial Neural Network	0.94

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

Figure 5: The ROC curve for the two models



4. Discussion

The purpose of this study is to build predictive models and select the one that has good performance, and figure out the correlations of the factors and children's chance for developing emotional and behavioral disorders. We built two models, the logistic regression and the artificial neutral network. And the two models both achieved a similar performance. Also, using Garson's algorithm, we are able to ascertain that the child's need for special care, special therapy, and weight are most related to emotional and behavioral disorders. Figure 2 also shows that these questions are important predictors of the dependent variable. Combining the results with Figure 1, we can see that in order to assess the child's chance of developing EBD, it will be most effective to look at factors such as the child's age and whether the child is currently receiving special care or special therapy.

According to the results of this study, parents, educators, and healthcare professionals can take appropriate measures to decrease their children's chance of developing EBD. For example, parents could strive to provide quality health care to children to ensure their mental health problems will not deteriorate into EBD. However, since most independent variables we selected for this report are innately determined, there's not much we can change. Thus, treatments are more important for teenager mental problems.

ISSN: 2455-8834

Volume:07, Issue:04 "April 2022"

One limitation of the study is the mean value imputation which we use the mean value to replace the missing value. This is a timesaving but flawed approach. We might have some new bias depending on the number of data that are imputed. For future studies, we can use more advanced techniques such as k-nearest neighbors (kNN) imputation, which replaces missing values with the mean of *k* (a parameter selected by the user) nearest neighbors of that sample. This technique requires more work but can generally achieve better performance and may help create a more accurate model.

References

CDC (2021). "Data and Statistics on Children's Mental Health.", Centers for Disease Control and Prevention, 22 Mar. 2021, www.cdc.gov/childrensmentalhealth/data.html.

Garcia-Arocena, D. (2015). *Happy or sad: The chemistry behind depression*. The Jackson Laboratory. https://www.jax.org/news-and-insights/jax-blog/2015/december/happy-or-sad-the-chemistry-behind-depression#.

Garson, G.D. 1991. Interpreting neural network connection weights. Artificial Intelligence Expert. 6(4):46-51.

Google. (2020 Aug. 11). *Classification: ROC curve and Auc | machine Learning crash course*. Google. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

National Survey of Children's Health - Data Resource Center for Child and Adolescent Health, 2019, www.childhealthdata.org/learn-about-the-nsch/NSCH.

Stefan Fritsch, Frauke Guenther and Marvin N. Wright (2019). neuralnet: Training of Neural Networks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet

Wang, Sijie, et al. "Exposure to Melamine cyanuric acid in adolescent mice caused emotional disorder and behavioral disorder." *Ecotoxicology and Environmental Safety* 211 (2021): 111938.

