Autism Disease Detection using Classification Algorithms

Md Tahmid Zoayed
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
2019-1-60-154@std.ewubd.edu

Plabon Banik
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
2019-1-60-167@std.ewubd.edu

Sayma Haque Arshe
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
2019-1-60-021@std.ewubd.edu

Mohammad Rifat Sarker
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
2019-1-60-169@std.ewubd.edu

Abstract

Autism is a neuro-disorder in which a person's interactions and communication with others are affected for the rest of their lives. Autism can be diagnosed at any age and is referred to as a "behavioral disease" since symptoms commonly develop in the first two years of life. Autism is a gradually growing concern among people of all ages. Early identification of this neurological condition can help keep the subject's mental and physical health in good shape. With the increased use of machine learning-based models to forecast numerous human diseases, early diagnosis based on multiple health and physiological parameters appears to be viable. This prompted us to become more interested in the diagnosis and investigation of ASD disorders in order to enhance treatment methods. We have used Classification algorithms of machine learning and tried to build the best model possible to detect Autism disease. We have taken the dataset from Kaggle to use our algorithms. We have made some data cleaning and pre-processing with our dataset, which has 706 rows and 22 columns with patient data. We have implemented 7 different algorithms on this dataset to get which algorithm suits best for this dataset. The algorithms are Decision Tree, Random Forest, K-Nearest-Neighbors, Naive Bayes, SVM, Logistic regression, and SGD. From different algorithms, we got different scores of efficiencies. After judging all of the scores, we found Random Forest the best fit for our dataset. This algorithm will help to detect Autism influenced patients without surgical operations. We can further implement this model as a GUI or software in different mental hospitals. This will reduce the expenses of an autism patient and make things much easier for hospitals as well.

Introduction

Autism is a neuro-disorder in which a person's interactions and communication with others are affected for the rest of their lives. Autism can be found at any age and is referred to as a "behavioral disease" since symptoms commonly develop in the first two years of life.

ASD detection is challenging since there are various different mental conditions with few symptoms that are quite similar to those with ASD symptoms, making this a difficult assignment. It's worth noting that both environmental and genetic factors could have a role in the development of this condition. The symptoms of this condition can begin as early as three years old and last a lifetime. Although it is impossible to completely heal patients suffering from this disease, the effects can be mitigated for a period of time if the signs are discovered early. The actual reasons for ASD have yet to be identified by scientists, who assume that human genes are to blame. Human genes influence development by altering the surrounding environment. There are certain risk factors that influence ASD, such as low birth weight, having an ASD sibling, and having elderly parents, among others. There are some social interaction problems they can suffering like as,

- Want to live alone
- Not able to make eye contact properly
- No interaction with others
- No sensitivity to pain
- May not have a wish for cuddling
- Inappropriate laughing and giggling
- No proper response to sound
- Inappropriate objects attachment etc.

Constrained interests and constant repetition of activities are often problematic for people with ASD. The list below includes specific instances of the many sorts of behaviors.

- Having a passing interest in specific aspects of the topic, such as figures and facts.
- Repeating specific habits
- In some circumstances, such as light, noise, and so on, you are less sensitive than another individual.
- When a person's routine is disrupted, they will feel upset.

Observation is generally used to identify ASD symptoms. ASD symptoms are typically detected by parents and instructors in older and adolescent students who attend school. Adults have a harder time diagnosing ASD symptom than older children and adolescents because certain ASD symptoms overlap with those of other mental health condition.

Autism is a rapidly growing concern among people of all ages. Early identification of this neurological condition can help keep the subject's mental and physical health in good shape. With the increased use of machine learning-based models to forecast numerous human diseases, early diagnosis based on multiple health and physiological parameters appears to be viable. Machine learning is now being used to diagnose disorders such as depression, ASD. The key goals of using machine learning techniques are to enhance diagnostic accuracy and minimize diagnosis time in order to enable faster access to health care services. Because the diagnosis of a case entails determining the correct class (ASD, No-ASD) based on the input case attributes, this process may be classified as a classification job in machine learning. In this study, we use a variety of classification algorithms to increase the accuracy of recognizing ASD patients across all four datasets. This prompted us to become more interested in the diagnosis and investigation of ASD disorders in order to enhance treatment methods.

Literature review

Several studies have made use of machine learning in various ways to improve and speed up the diagnosis of ASD. Duda et al. [5] applied forward feature selection coupled with under sampling to differentiate between autism and ADHD with the help of a Social Responsiveness Scale containing 65 items. Deshpande et al. [4] used metrics based on brain activity to predict ASD. Soft computing techniques such as probabilistic reasoning, artificial neural networks (ANN), and classifier combination have also been used [15]. Many of the studies performed have talked of automated ML models which only depend on characteristics as input features. A few studies relied on data from brain neuroimaging as well. In the ABIDE database, Li et al. [14], extracted 6 personal characteristics from 851 subjects and performed the implementation of a cross-validation strategy for the training and testing of the ML models.

Vaishali R, Sasikala R. et al. [3] have proposed a method to identify Autism with optimum behavior sets. In this work, an ASD diagnosis dataset with 21 features obtained from the UCI machine learning repository experimented with swarm intelligence based binary firefly feature selection wrapper. The alternative hypothesis of the experiment claims that it is possible for a machine learning model to achieve a better classification accuracy with minimum feature subsets. Using Swarm intelligence based single-objective binary firefly feature selection wrapper it is found that 10 features among 21 features of ASD dataset are sufficient to distinguish between ASD and non-ASD patients. The results obtained with this approach justifies the hypothesis by producing an average accuracy in the range of 92.12%-97.95% with optimum feature subsets which are approximately equal to the average accuracy produced by the entire ASD diagnosis dataset.

Li B, A. Sharma, J Meng, S. Purushwalkam, E. Gowen (2017) et al. [11] have used machine learning classifiers to detect autistic adults by imitation method. The goal of this study was to investigate the fundamental problem related to discriminative test conditions and kinematic parameters. The dataset contains 16 ASC participants who have a series of hand movements. In this 40 kinematic constraints from 08 imitation conditions has been extracted by using machine learning methods. This research shows that for a small sample, there is a feasibility of applying machine learning methods to analyze high-dimensional data and the diagnostic classification of autism. The sensitivity rates achieved by RIPPER which have the features Va (87.30 %), CHI (80.95%), IG (80.95%), Correlation (84.13%), CFS (84.13%), and “no feature selection”(80.00%) on the AQ-Adolescent dataset.

It is evident from the above discussed section that there is definitely a need to explore the possibility of applying deep learning based models for the detection of ASD in human population. Most of the work discussed above use conventional machine learning approaches and hence are limited in their performance. In this work, performance of several machine learning models have been compared to that of the deep learning model for this purpose. Separate models have been prepared for separate population set (discussed in section below) and compared individually.

DATASET DESCRIPTION

The dataset for this study was obtained from Kaggle, which is open to the public. This autism screening of adult’s dataset has 706 records of ASD patients with 20 attributes (features). These features are divided into two types, behavioural features, and distinctive features. It tracks ten (AQ-10-Adult) as well as ten individual characteristics of the patients. Here the attribute types are mainly Categorical, continuous, and binary. There is also a classification attribute(class).

List of Attributes in the dataset

Attribute Id	Attribute description
(1-10)	AQ scores
11	patients Age
12	patients Gender
13	Ethinicity
14	The patient suffered from Jaundice problem by birth
15	autism
16	The country in which the user lives
17	Screening Application used by the user before or not?
18	result

19	age description
20	relation

To the best of our knowledge, this dataset was used for identifying ASD cases.

Dataset Cleaning

The data has different types of values; some are categorical, some are numerical and some of them are in strings. We have cleaned the dataset in such a way that we find a close & appropriate result. We checked the correlation of the columns with the prediction column and tried to check the correlations. The co-relations which are closer to zero were the unnecessary columns. 'used_app before', 'id'.'age_desc' are those columns which are dropped. We have also got some nan values marked as '?'. We have also filtered them out. We have also looked at our data set using description and info functions. There are some string values in our dataset; we have changed their types. We have also made a pipeline using Simple Imputer to clean data which are yet to add. This is kind of an automation system.

Description of used algorithms

We have used 7 different types of algorithms in this dataset to find different results generated from this dataset. They are given below with a short description

Decision Tree

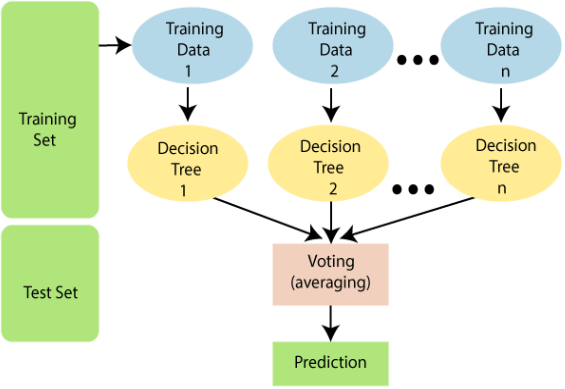
Decision Tree may be a Directed learning procedure that can be utilized for both classification and Relapse issues, but for the most part, it is preferred for understanding Classification issues. It may be a tree-structured classifier, where inner hubs speak to the highlights of a dataset, branches speak to the choice rules and each leaf hub speaks to the result. In call Trees, for predicting a category label for a record we tend to begin from the basis of the tree. we tend to compare the values of the basic attribute with the record's attribute. On the idea of comparison, we tend to follow the branch reminiscent of that price and jump to the following node.
Types of Decision Trees:

1. Categorical Variable Decision Tree
2. Continuous Variable Decision Tree

Random Forest

Random Forest is a well-known machine learning calculation that has a place in directed learning Strategy. It can be utilized for Classification and Regression issues as well in Machine Learning. It is based on the concept of gathering learning, which could be a handle of combining numerous classifiers to unravel a complex issue and to progress the execution of the demonstration.
Properties of random forest are given below:

1. It takes less preparation time as compared to other calculations.
2. It predicts yield with tall precision, indeed for the expansive dataset, it runs effectively.
3. It can keep up exactness when an expansive extent of information is lost.



K-Nearest-Neighbour

The k-nearest neighbors (KNN) calculation could be a basic, easy-to-implement administered machine learning calculation that can be utilized to illuminate both classification and relapse issues. The k-nearest neighbors (KNN) calculation could be a basic, easy-to-implement administered machine learning calculation that can be utilized to illuminate both classification and relapse issues. "Lazy learner" algorithms, such as the k-nearest-neighbor, don't create a model of the data set before using it. Calculations are only performed if the data point's neighbors are asked to be polled. This makes k-nn a breeze to use in data mining applications.

Support vector machine

A support vector machine (SVM) could be an administered machine shows that employments classification calculations for two-group classification issues. After giving an SVM demonstrate sets of labeled training data for each category, they're able to classify unused content. To be sure, it's most commonly employed in the context of classifying data. According to this method (where n is the number of features you have), we plot each data item as an individual point in space with the value of each feature being a coordinate.

Naive Bayes

The naive Bayes classifier could be a generative demonstration for classification. Sometime recently the appearance of profound learning and its easy-to-use libraries, the Naive Bayes classifier was one of the broadly sent classifiers for machine learning applications. Despite its straightforwardness, the naive Bayes classifier performs very well in many applications. As a result of this assumption, the model is said to as naïve. In other words, changing the value of one feature has no direct impact on or effect on other characteristics utilized in the algorithm, and vice versa. Alright. There's no doubt that Naive Bayes is an efficient and straightforward method. But why has it become so popular? The reason for this is that New Brunswick has a substantial edge. We can easily write the method and make predictions because it's a probabilistic model.

Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) could be a straightforward however exceptionally productive approach to fitting direct classifiers and regressors beneath curved misfortune capacities such as (direct) Bolster Vector Machines and Calculated Regression. Even though SGD has been around within the machine learning community for a long time, it has gotten an impressive sum of consideration fair as of late within the setting of large-scale learning.

Logistic regression

In the Supervised Learning approach, one of the most use-full Machine Learning algorithms is logistic regression. It's a method for predicting a categorical dependent variable from a set of independent variables. It's important to understand the following assumptions regarding logistic regression before going into its implementation: It's important to note that when using binary logistic regression the target variables must always be binary. The intended outcome is indicated by factor level 1. As a result, the independent variables in the model must be independent of each other. Meaningful variables must be included as part of the analysis process. For logistic regression, we need to use a big sample size.

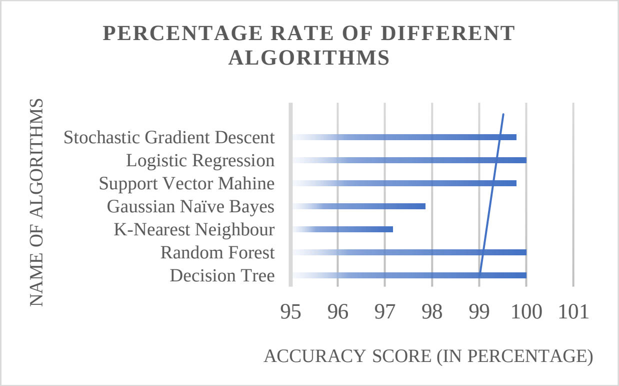
Result & Analysis

We have implemented 7 different types of algorithms on this data set and analyzed the percentage of the score for different types of algorithms. The deciding factor in this data set is "Class/ASD" column which had two values," Yes" and 'No'. "Yes" means the person has autism and 'No' means he/she is fit. We have got different results for different algorithms and we have chosen the best last Decision tree, Random Forest, K- Nearest Neighbor, Gaussian Naive byes, logistic regression, and SGD are the algorithms we have used in this data set. Among them, Random Forest, Logistic Regression, and Decision tree algorithm give the best result. The result of other algorithms and their comparison is shown down using a table and bar plot.

Name of Algorithms	**Accuracy Score (in Percentage)**
Decision Tree	100
Random Forest	100
K-Nearest Neighbor	97.165
Gaussian Naïve Bayes	97.86
Support Vector Machine	99.79
Logistic Regression	100.0
Stochastic Gradient Descent	99.79

Table on accuracy score of different Algorithms

From the table, we can see K-Nearest-Neighbor and Gaussian Naïve Bayes gives the lowest which is 99.7% and Random Forest, Logistic Regression and Decision tree give the highest score of 100%. If we visualize them in a bar chart then they would look like this.



Bar chart on accuracy score of different Algorithms

After getting the result we have made a joblib function called 'autism.joblib' to call and use the model for giving it a GUI or using it as a software. We also did a test using the 'autism.joblib' and get our desired result.

Conclusion

In this paper, different machine learning and deep learning approaches were used to detect ASD. We will keep the machine learning technique with optimal results. We'll try different algorithms and finally select the best performing algorithm for this dataset which will help us to get a better version of the result. Random forest and logistic regression-based models show the best accuracy of prediction for the ASD dataset. Using machine-learning technologies to improve diagnosis, and support for people on the spectrum. ASD diagnosis can significantly reduce the neurodevelopment problem in our society.

REFERENCES

[1] Fadi Thabtah. (2017) “ASD Tests. A mobile app for ASD screening.” www.asdtests.com [accessed December 20th, 2017].

[2] Baihua Li, Arjun Sharma, James Meng, Senthil Purushwalkam, and Emma Gowen. (2017) “Applying machine learning to identify autistic adults using imitation: An exploratory study.” PloS one, 12(8): e0182652.

[3] Fadi Fayez Thabtah (2017), “Autistic Spectrum Disorder Screening Data for Adult”., https://archive.ics.uci.edu/ml/machine-learning-databases/00426/.

[4] M. S. Mythili, and AR Mohamed Shanavas. (2014) “A study on Autism spectrum disorders using classification techniques.” International Journal of Soft Computing and Engineering (IJSCE), 4: 88-91.

[5] J. A. Kosmicki, V. Sochat, M. Duda, and D. P. Wall. (2015) “Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning.“ Translational psychiatry, 5(2): e514.

[6] Fadi Fayez Thabtah (2017), “Autistic Spectrum Disorder Screening Data for children,” https://archive.ics.uci.edu/ml/machine-learning-databases/00419/ ,2017

[7] Fadi Fayez Thabtah (2017), “Autistic Spectrum Disorder Screening Data for Adolescent”, https://archive.ics.uci.edu/ml/machine-learning-databases/00420/.

[8] John, George H., and Pat Langley. (1995). “Estimating continuous distributions in Bayesian classifiers.” In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (pp. 338-345). Morgan Kaufmann Publishers Inc.

[9] Quinlan, J. R. (1993) “Program for machine learning.” C4. 5.

[10] Keerthi, S. Sathiya, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and Karuturi Radha Krishna Murthy. (2001) “Improvements to Platt's

1004 Suman Raj et al. / Procedia Computer Science 167 (2020) 994–1004 Author name / Procedia Computer Science 00 (2019) 000–000 11 SMO algorithm for SVM classifier design. “ Neural computation, 13(3):637-649.

[11] Ra j, S., Masood, S.: Analysis and detection of autism spectrum disorder using machine learning techniques.Procedia Computer Science 167, 994 – 1004 (2020)

[12] Thabtah, F.: Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. dl.acm.org Part F129311, 1–6 (2017)

[13] Thabtah, F.: ASD Tests. A mobile app for ASD screening. (accessed January 10, 2020). URL https://www.

[14] Thabtah, F.: ASD Dataset (accessed January 13, 2020). URL https://fadifayez.com/autism-datasets/

[15] Thabtah, F.: ASD Dataset- UCI machine learning repository, 2017 (accessed January 13, 2020). URL https:

[16] Thabtah, F., Abdelhamid, N., Peebles, D.: A machine learning autism classification based on logistic regression analysis. Health Information Science and Systems 7(1), 1–11 (2019)

[17] Thabtah, F., Peebles, D.: A new machine learning model based on induction of rules for autism detection. Health informatics journal 26(1), 264–286 (2020)

[18] Wiggins, L.D., Reynolds, A., Rice, C.E., Moody, E.J., Bernal, P., Blaskey, L., Rosenberg, S.A., Lee, L.C., Levy, S.E.: Using Standardized Diagnostic Instruments to Classify Children with Autism in the Study to Explore Early Development. Journal of Autism and Developmental Disorders 45(5), 1271–1280 (2015)