|  |
| --- |
|  |

|  |
| --- |
| DSCI- 6001-01 – Intro to Data Science |

|  |  |
| --- | --- |
| FALL 22 |  |



Contents

[Executive Summary 2](#_Toc121347075)

[Highlights of Project: 4](#_Toc121347076)

[Submitted on: Dec 8 2022 4](#_Toc121347077)

[Technical Report 4](#_Toc121347078)

[Abstract: 5](#_Toc121347079)

[Methodology 10](#_Toc121347080)

[Results & Discussion 24](#_Toc121347081)

[Conclusion 27](#_Toc121347082)

[Contributions/References 28](#_Toc121347083)

|  |
| --- |
| **A Machine Learning Approach to Predict Autism Spectrum Disorder** |

|  |  |  |
| --- | --- | --- |
| Executive Summary  * PROBLEM STATEMENT   "To design and implement a system, which will be able to detect autism at quite an early stage using machine learning algorithms"  Input: AQ-10 dataset and certain other attributes such as age, gender, etc.  Output: Identifies whether the individual should be referred for a comprehensive autism assessment.   * PROPOSED SYSTEM   The objective of this work is to propose an autism prediction model using ML techniques that could Effectively predict autism traits of an individual of any age.  The user can answer the questions and based on the fitted model it will be predicted whether the person has autism or not.   * AIM OF THE PROJECT   Earlier detection of autism can come to a great help by prescribing patients with proper medication at an early stage. Thus, a time-efficient, accurate and easy screening test tool is very much required which would predict  Autism traits in an individual and identify whether they require comprehensive autism assessment.   * OBJECTIVES OF THE PROJECT  1. To design a system that can help detect autism at an early stage. 2. To minimize the overfitting problem introduced in simple ID3 algorithm. 3. To increase the accuracy of the system compared to existing ones. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**   1. **Setranah Selinam** 2. **Devnath Reddy Motati** 3. **Ajay Kumar Jagu** 4. **Avinash Reddy Bhavanam** 5. **Krishna Surendra Palepu** | **Questions?**  Contact : ssetranah@gmail.com |  |

**A Machine Learning Approach to Predict Autism Spectrum Disorder**

## Highlights of Project:

\* Collected Data from Kaggle.

\* Find correlation between the data & dropped least correlated data.

\* Prepared the data according to requirement using Jupiter Notebook.

\* Model the Data Using Random Forest, Adaboost, Logistic Regression, SVM Classifier Algorithms.

\* Compared accuracies of all models & choose best of them.

\* Deployed using Flask.

## Submitted on: Dec 8 2022

|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
|  |  |
|  |

## Abstract:

In present day autism spectrum disorder (ASD) is gaining its momentum faster than ever. Detecting autism traits through screening tests is very expensive and time consuming. With the advancement of artificial intelligence and machine learning (ML), autism can be predicted at quite early stage. Though number of studies have been carried out using different techniques, these studies didn’t provide any definitive conclusion about predicting autism traits in terms of different age groups. The proposed model was evaluated with AQ- 10 dataset and 1000 real dataset collected from people with and without autistic traits. The evaluation results showed that the proposed prediction model provide better results in terms of accuracy, specificity, sensitivity, precision and false positive rate (FPR) for both kinds of datasets.

Introduction

Autism spectrum disorder is a neuro-developmental disorder that affects a person’s interaction, Communication and learning skills. People with autism might have problems with learning disabilities, mental Health problems such as anxiety, depression etc, motor difficulties, sensory problems. For example, they could Have trouble communicating but be unusually good at art, music, math, or memory. Because of this, they Might do especially well on tests of analysis or problem-solving. The common symptoms that a autistic person shows are lack of eye contact, a narrow range of interests or intense interest in certain topics, Doing Something over and over, like repeating words or phrases, rocking back and forth, or flipping a lever, High Sensitivity to sounds, touches, smells, or sights that seem ordinary to other people, Not looking at or listening To other people, etc. Autism treatment is costly at later stages. So, if autism gets detected at an early age, then Treatment is easy and time and money can be saved. Although the diagnosis of autism can be done at any age, its Symptoms generally appear in the first two years of life and develop through time.[1] Autism spectrum disorders (ASD) are a diverse group of conditions. They are characterized by some degree of Difficulty with social interaction and communication.[2] Other characteristics are atypical patterns of activities and behaviors, such as difficulty with the transition from one activity to another, a focus on details, and unusual Reactions to sensations.[2] The abilities and needs of autistic people vary and can evolve over time. While some people with autism can live independently, others have severe disabilities and require life-long care and support. Autism often has an impact on education and employment opportunities.[3] In addition, the demands on families to provide care and support can be significant. Societal attitudes and the level of support provided by local and national Authorities are important factors determining the quality of life of people with autism.[3] The Diagnosis takes many hours [8], and the continuously growing demand for appointments is much greater than The peak capacity of the country’s pediatric clinics [8], Detecting and treating autism spectrum disorder in its Early stages are extremely crucial as this helps to decrease or alleviate the symptoms to a certain extent, thus Improving the overall quality of life for the individual. However, owing to the gaps between initial concern and diagnosis, a lot of valuable time is lost as this disorder remains undetected. Machine Learning methods Would not only help to assess the risk for ASD in a quick and accurate manner but are also essential to Streamline the whole diagnosis process and help families access the much-needed therapies faster. Some of The screening methods used to detect ASD are Autism Spectrum Quotient (AQ). In this project, AQ10 is used as a screening method .[9] Most parents report that the onset of autism features appears within the first year of life. There are two possible Developmental courses for ASD. One course of development is more gradual in nature, which parents report Concerns in development over the first two years of life and diagnosis is made around 3-4 years of age. Some of the early signs of ASDs in this course include decreased attention at faces, failure to obviously respond When name is called, failure to show interests by showing or pointing, and delayed imaginative play. [9] A second course of development is characterized by normal or near-normal development before onset of Regression or loss of skills. One pattern of regression occurs in the first 1 5 months to 3 years. There are some Who believe that regressive autism is simply early-onset autism that was recognized at a later date. Researchers have conducted studies to determine whether regressive autism is a distinct subset of ASD [7]. Over the years, the results of these studies have contradicted one another. Some researchers believe there is Still nothing to support a definitive biological difference between early-onset and regressive autism.

Thus, a time efficient, accurate and easy screening test tool is very much required which would predict autism Traits in an individual and identify whether or not they require comprehensive autism assessment

Review of available research

[1] A Machine Learning Approach to Predict Autism Spectrum Disorder Kazi Shahrukh Omar, Prodipta Mondal, Nabila Shahnaz Khan, Md. Rezaul Karim Rizvi, Md Nazrul Islam Year-2019

METHODOLOGY:

At first Decision Tree-CART algorithm was implemented to predict autism traits in an individual. For further improvement Random Forest-CART was implemented and better results were obtained. Finally, Random Forest-CART classifier was modified to get improved results by merging it along With Random Forest-ID3 classifier.

LIMITATIONS:

Lack of sufficiently large data to train the prediction model. Not designed for age group below 3 years.

[2] “Automatic Detection and Labeling of Self-Stimulatory Behavioral Patterns in Children with Autism Spectrum Disorder Cheol-Hong Min Year-2017

METHODOLOGY:

The system incorporates 2 different sensor platforms-the wearable system is which detects behavioral Patterns of a subject, while the static sensors are captures the sounds, images and videos of the subjects Within a room. The system stores the time of the video when the activities are detected. Time-Frequency methods are used to extract features and Hidden Markov Model (HMM) are used for analyzing the accelerometer signal.

LIMITATIONS:

Misses instances such as single flapping or single tapping due to short duration of the events.

[3] “Detecting Autism Based on Eye-Tracking Data from WebSearching Tasks” Victoria Yaneva, Le AnHa, Sukru Eraslan, Yeliz Yesilada,Ruslan Mitkov Year-2018

METHODOLOGY:

This paper is based on the atypical visual-attention patterns of people with autism.

Gaze data is collected from two different kinds of task-Browsing and Searching which is then used to train a machine learning classifier. The classifier is based on logistic regression.

LIMITATIONS:

The above method doesn't perform well for images with more varied and extreme transformations especially geometric changes.

There is still a lingering gap between the results achievable with paired training data and those achieved by our unpaired method.

[4] Enhancing Diagnosis of Autism with Optimized Machine Learning Models and Personal Characteristic Data Milan N, Parikhl, Hailong Li and Lili Hel Year-2019

METHODOLOGY:

Nine automated machine learning models were developed by using a large PCD (Personal Characteristic data) dataset from the ABIDE repository.The six PCD features selected were- age, sex, handedness, fiull-scale IQ, verbal IQ, performance IQ

Out of all the models, the neural network model performed the best followed by k-nearest neighbor.

LIMITATIONS:

Exaggeration is more obvious in the face shape than facial features due to corresponding landmarks being the principal components. Content and style entangled in geometry.

## Methodology

The project involved CRISP methodology, which includes:

* Business understanding
* Data understanding
* Data preparation
* Modelling
* Evaluation
* Deployment

Diagram

Description automatically generated

MODULE DESCRIPTION

GATHERING DATA

Here data from Kaggle is used. It is based on 10 screening questions focus on different domains such as­ attention to detail, attention switching, communication, imagination and social interaction. For each question, it is either 1 or 0. It also contains certain features such as Age, Gender, if born with Jaundice.

* + - Dataset Description: To apply supervised algorithms there is a usage 1054 datasets. It has been categorized into three areas including medical, health and social science. Besides, the attribute type is categorical, continuous, and binary. The queries of the Q-CHAT and AQ tools are 10. Each item value assigned from Kaggle and UCI ML repository [4]. The contained datasets represent 30.76% female and 69.24% male. There are a total of 1054 toddler case values. In addition, the following ten questions were asked to the parent, self, caregiver, and medical staff. Nevertheless, the table below describe the feature of data in different columns [4].

|  |  |
| --- | --- |
| Variables in Dataset | Corresponding features |
| Al | I often notice small sounds when others do  not |
| A2 | I usually concentrate more on the whole  picture, rather than the small details |
| A3 | I find it easy to do more than one thing at  once |
| A4 | If there is an interruption, I can switch back to  what I was doing very quickly |
| A5 | I find it easy to 'read between the lines' when  someone is talking to me |
| A6 | I know how to tell if someone listening to me  is getting bored |
| A7 | When I'm reading a story L find it difficult to  work out the characters' intentions |
| A8 | I like to collect information about categories of  things (e.g. types of car, types of bird, types  of train, types of plant etc) |
| A9 | I find it easy to work out what someone is  thinking or feeling just by looking at their face |
| A10 | I find it difficult to work out people's intentions |

PREPARING THE DATA

The collected data is synthesized to remove irrelevant features. Observations with null columns are removed. And string values are converted to integer values.

* + - Data pre-processing: The acquired dataset needed some modifications before it could be tested on the classifier algorithms. Pre-processing the dataset in such a way that it was able to provide prime output. Previously it was found that unsorted and unprocessed data affected our result scores hugely. To get rid of the problems some particular steps were followed so the algorithms would be able to give more precise results. Most of the values in the dataset were binary and Boolean values. They were based on polar questions primarily. However, few of the question criteria required non integer and non-Boolean type answers. These data were recorded in string format. For the algorithm to give the optimal result needed to convert these string type values to binary, used one hot encoding to transform the values to binary. One hot encoding is a method that converts categorical variables into a type which can be given to ML algorithms to do a better job of projection. For this process first need to switch the string values of column by default then applied one hot encoding on these values to get the unique binary correspondents.

The ML algorithms which will be used for experimentation based on the data size, the data features and the target set of results are looking for. Initially the following algorithms were chosen:

* + - * Random forest Classifier
      * SVM Classifier
      * Logistic Regression
      * Adaboost Classifier

For these algorithms there is a need to import the required libraries individually for each of the classifiers. The goal by applying these classifiers was to find the accuracy, precision and recall scores for which the necessary libraries were imported. For Random Forest and AdaBoost classifier along with the other required ones. Some of the common libraries all the libraries used were pandas, NumPy, plt, plot. The experiments were done using Spyder and all libraries were imported, and the pre-processed data was deployed on Spyder to compute the target result. At the beginning of the process the features of the data were set to X and the output to Y. Then splatted data was set into a test and trainset, where 20% of the dataset were used for testing and 80% were used for training.

CHOOSING A MODEL

A machine learning classifier is trained using 80% of the training data. The classifier is based on Random Forest.

The proposed model ensures a more accurate result to the research about autism detection at an early age of autism. This project used machine learning algorithms to get more accurate results like the Random Forest. First, the questions were collected by following a standard. Those who are required to answer them the values were collected from their answers and sent them to data pre-processing. Data pre-processing is adjusting if it is a non-matrix format type or not, how many attributes and instances are there and how many need to run in a specific algorithm.

Diagram

Description automatically generated

The collected data of question sets the train our machine learning algorithm. For that the necessary libraries were imported. After that allocated 80% of the dataset was set for training and 20% for testing. The X matrix for feature and Y matrix for output. By using one hot encoding that transforms all 'AQl0' columns to unique binary values. Accuracy of the algorithms has been compared to each other for better results. Furthermore, the score for precision and recall were formulated. Precision does not depend on accuracy as for precision the values are not always the same. After getting precision, accuracy and the results were compared and the best fit algorithm for the model was selected. Lastly the selected algorithm was applied for the model. The goal is to improve the accuracy of the result differing from the other autism detection results.

EVALUATION

The Pareto Principle is also called the 80/20 rule. The general point is that, in most cases, 80% of effects come from 20% of causes.

The model is evaluated using Pareto principle which composes of 20% of the original data. By using Confusion matrix, the accuracy of the model can be determined. Now a confusion matrix which basically is for defining the results of a classification algorithm. Accuracy, Precision, Recall, Fl score can be calculated from the left most four values. And these four values or parameters are true negative, true positive, false negative and false positive which are the answers of class output. In the following table the true positives and true negatives are shown in green color because these are correctly predicted. Also, false positives and false negatives are shown in red color because these are incorrectly predicted. These red color values are needed to be minimized.

Table

Description automatically generated

The details of these values are: -

* + - True Positives (TP): True positives are correctly found values. This means the genuine result and the anticipated result both are yes.
    - True Negatives (TN): True negatives are also correctly found results. This means the genuine and anticipated result both are no.
    - False Positives (FP): This means the real result is no, but the anticipated result is yes
    - False Negatives (FN): This means the real result is yes, but anticipated result is no.

To identify how good a model has performed can be measured by some parameters namely Precision, Accuracy, Fl Score and Recall. By using these four values to calculate Precision, Accuracy, Fl Score and Recall. These scores are:

* Accuracy: Accuracy means how close the measured value is compared to the standard value. Proportion of the total number of predictions that are correct.

Accuracy = (T P + T N) / (T P + F P + F N + T N)

* + - Precision: It indicates to the adjacency of two or more measurements to each other. If it takes any measurement 3 times and every time if there is a same value though the value is not close to the standard value then the measured value will be considered as precise but not accurate. So precision does not depend on accuracy.

Precision = TP / (T P + F P)

* + - Recall: Recall actually detentions how many of the real positives a model captures by labelling it as true positive. So, recall is accurately anticipated positive inspection ratio to all the inspections in the real table.

Recall = TP / (T P + F N)

* + - Fl Score: F1 takes both false negative and false positive in the count and takes a weighted average. F1 is usually more useful than accuracy.

FlScore = 2 \*((Precision\* Recall)/ (Precision+ Recall))

PREDICTION

The user will be entering the real time data in the AS Detect web application and based on the answers to the AQ-10 questions and other parameters the pie chart representation is used to predict whether the person has autism or not.

SYSTEM ARCHITECTURE

System architecture is the conceptual design that defines the structure and behavior of a system. An architecture description is a formal description of a system, organized in a way that supports reasoning about the structural properties of the system. It defines the system components or building blocks and provides a plan from which

products can be procured, and systems developed, that will work together to implement the overall system.

Diagram

Description automatically generated

USE CASE DIAGRAM OF THE SYSTEM

A use case diagram is a type of behavioral diagram created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases.

Diagram

Description automatically generated

Diagram

Description automatically generated

Models Used:

1. RANDOM FOREST:

Chart, diagram, radar chart

Description automatically generated

* + It is an ensemble method for classification, regression and other task that operates by constructing a multitude of decision trees at training time.
  + When training, each tree in a random forest learns from a random sample of the data points. The samples are drawn with replacement, known as bootstrapping, which means that some samples will be used multiple times in a single tree.
* At test time, predictions are made by averaging the predictions of each decision tree m case of regression. It considers the maximum counted prediction in case of classification

#### SVM:

Chart, scatter chart

Description automatically generated

* + A support vector machine (SVM) is a supervised [machine learning](https://monkeylearn.com/machine-learning/) model that uses [classification algorithms](https://monkeylearn.com/blog/machine-learning-algorithms/) for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they’re able to categorize new text.
  + Compared to newer algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands).
  + This makes the algorithm very suitable for text classification problems, where it’s common to have access to a dataset of at most a couple of thousands of tagged samples.

#### ADBOOST:

* + AdaBoost is a ML Algorithms that is adaptive in which subsequent weak learners are trained in favour of those instances misclassified by previous classifiers.
  + It is best used to boost the performance of decision trees on binary classification problems.
  + Each instance in the training dataset is weighted. The initial weight is set to: weight(xi) = 1/n, where xi is the ith training instance and n is the number of training instances.
  + These instances are input to one of the models. The weight of the instance wrongly classified is reduced and again the instances are given as the input to the next model and so on. In this way the predictions are made

#### LOGISTIC REGRESSION:

* + In statistics, the logistic model is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables.
  + In regression analysis, logistic regression is estimating the parameters of a logistic model

## Results & Discussion

Evaluation results of Random Forest Model

Text

Description automatically generated

Evaluation Results of SVM Model

Graphical user interface, text, application

Description automatically generated

Comparison of Models:

A black screen with white text

Description automatically generated with medium confidence

From our data frame of accuracies and f1\_scores from various models, we realize that the Random Forest and Logistic Regression models give the best accuracies

we decided on Random Forest is best for the predictions.

Based on comparison of Logistic, SVM, Adaboost, Random Forest Model we came to decision to use Random Forest because it gives higher accuracy and it is best for Predictions.

WEB Deployed Outcome:

Graphical user interface, text, application, email

Description automatically generated

Here we get the results based on User input for Questions

It shows positive if the user has autism otherwise it shows negative.

The model can be accessed via <http://127.0.0.1:5000/index>

And all technical aspects of our project including jupyter notebook for data analysis and html, css, python codes for our deployment in our github via <https://github.com/sortt-outt>

## Conclusion

A prediction model was developed to predict autism traits. Using the AQ-10 data set, the proposed model can predict autism with more accuracy. This result showed better performance comparing to the other existing approach of screening autism.

A user-friendly application has been developed for end users based on the proposed prediction model so that any individual can use the application to predict the autism traits easily. This outcome indicated an extension of many other existing works, since most of the existing works mainly focus on developing and comparing the performance of prediction model or techniques and did not expend to develop any mobile application for end

users.

With the help of autism screening application, an individual can be guided at an early stage that will prevent the situation from getting any worse and reduce costs associated with delayed diagnosis

## Contributions/References

[1] Automatic detection and labeling of self-stimulatory behavioral patterns in children with Autism Spectrum Disorder Cheol-Hong Min 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)

[2] A Machine Learning Approach to Predict Autism Spectrum Disorder Kazi Shahrukh Omar; Prodipta Mondal ; Nabila Shahnaz Khan ; Md. Rezaul Karim Rizvi ; Md Nazrul Islam,2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)

[3] Machine learning for early detection of autism (and other conditions) using a parental questionnaire and home video screening,Halim Abbas ; Ford Garberson ; Eric Glover ; Dennis P. Wall,2017 IEEE Interational Conference on Big Data (Big Data)

[4] Autism Barta - A smart device-based automated autism screening tool for Bangladesh Sharmistha Bardhan; G. M. Monjur Morshed Mridha ; Eshtiak Ahmed ; M. Anwar Ullah Helal Uddin Ahmed ; Shaheen Akhter ; Md. Golam Rabbani ; Khondaker Abdullah Al Mamun,2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)

[5] "SenseA"-Autism Early Signs and Pre-Aggressive Detector through Image Processing Chanuki Gamaethige ; Umaya Gunathilake; Dhimanshi Jayasena ; Hansani Manike ; Pradeepa Samarasinghe ; Thilini Yatanwala 2017 Asia Modelling Symposium (AMS)

[6] Mamun, K. A. A., Bardhan, S., Ullah, M. A., Anagnostou, E., Brian, ]., Akhter, S., & Rabbani, M. G. (2016). Smart autism - A mobile, interactive and integrated framework for screening and confirmation of autism. 20 16 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

[7] Yaneva, V, Ha, L. A. Eraslan, S., Yesilada, Y., & Mitkov, R. (2018). Detecting Autism Based on Eye-Tracking Data from Web Searching Tasks. Proceedings of the Internet of Accessible Things on - W4A '18

[8] Parikh, M. N., Li, H., & He, L. (2019). Enhancing Diagnosis of Autism With Optimized Machine Learning Models and Personal Characteristic Data. Frontiers in Computational Neuroscience, 13.

[9] Identifying Children with Autism Spectrum Disorder Based on Their Face Processing Abnormality: A Machine Learning Framework” WenboLiu, Ming Li, and Li Yi Year2016