Early Diagnosis of Autism

Rahul Torlapati, Grusha Dahiya and Stephanie Richeson
rahult1@umbc.edu, grushad1@umbc.edu, sriches1umbc.edu
University of Maryland, Baltimore County

***Abstract* - Autistic Spectrum Disorder (ASD) is a neurodevelopmental condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Although autism can be diagnosed at any age, it is said to be a "developmental disorder" because symptoms generally appear in the first two years of life. ASD occurs in all ethnic, racial, and economic groups. Although ASD can be a lifelong disorder, treatments and services can improve a person's symptoms and ability to function. While scientists don't know the exact causes of ASD, research suggests that genes can act together with influences from the environment to affect development in ways that lead to ASD.**

INTRODUCTION

Autism spectrum disorder (ASD) affects approximately one in 59 children in the United States (U.S.). In 2007, the American Academy of Pediatrics (AAP) issued a strong recommendation for all primary care providers to screen children for autism during a well-child visit, to begin the referral process for more formal testing, and intervention. But this hasn't helped in identifying ASD cases more rapidly. So, we aim to create a few Machine Learning models which can accurately predict positive and negative ASD cases based the data gathered from AQ-10 test which is currently the standard testing procedure for people who are suspected to have ASD but do not have a learning disability. AQ-10 test is developed by University of Cambridge in association with Autism research center and is implemented across the U.S by the National Institute for Health Research. This test is recommended in 'Autism: recognition, referral, diagnosis and management of adults on the Autism Spectrum by NICE clinical guideline GC142. Although the test only covers a smaller portion of people who are affected with ASD, we believe working on this is a step in the right direction.

The AQ-10 test consists of 10 questions and 4 relative answers ranging from Definitely agree to Definitely disagree. The questions considered in the test are

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

Scoring criteria for this test is only 1 point can be scored for each with a maximum of 10 points. A person is scored 1 point for Definitely or Slightly agree on each of items 1, 7, 8 and 10. Person is scored 1 point for Definitely or Slightly disagree on each of items 2, 3, 4, 5, 6 and 9. If the individual scores more that 6 out of 10, he/she is considered for a Specialist Diagnostic Assessment.

We are using the data collected and curated by Thabtah, F. (2017), Department of Digital Technology at Manukau Institute of Technology in Auckland, New Zealand. The dataset is available on the UCI's Machine Learning Repository. Additionally, we are using a dataset which consists of curated information of 1059 toddlers from July of 2018. This dataset also consists of information as to who filled the AQ-10 on behalf of the child. Both the datasets are of non-matrix format type and consists of Categorical, continuous and binary data.

TOOLS USED

We used Python3 as the base coding language and Jupyter Notebook as the editor and environment. In Python3 we used libraries like Metrics, Pandas, Sklearn, NumPy, Seaborn, Matplotlib,

EXPLORATORY DATA ANALYSIS

Considering that fact that the dataset consists of information whether as to person/toddler is born with Jaundice we initially assumed that Jaundice at birth could be one of the deciding factors as whether the person/Toddler develops ASD in the future. But as it turned out Jaundice at Birth did not have any correlation to the development of spectrum disorder.

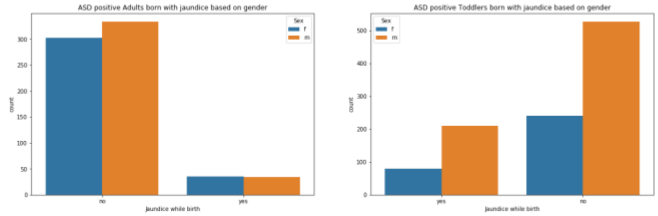

FIGURE I
COMPARISON BETWEEN ADULTS AND TODDLERS BORN WITH JAUNDICE BASED ON GENDER

We can see here almost 6-7 times more (in Adults) and 2-3 times more (in Toddlers) of non-jaundice born are ASD positive whereas according to reports that is around 10 times. Jaundice born child has a weak link with ASD. Also, according to reports, ASD is more common among boys (around 4-5 times) than among girls. But here in Adults we see a lower ratio, whereas in Toddlers its nearly 4 times boys than girls, which is quite close to actual ratio.

While researching on AQ-10 test, we came across several question as to whether the questions had relevance in determining ASD positiveness. We used a correlation Map of the dataset to determine the relevance of the AQ-10 to positiveness of ASD.

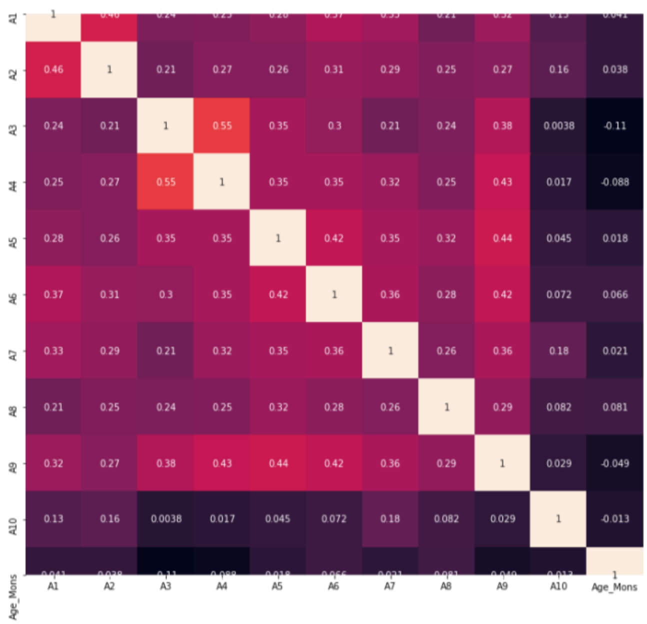

FIGURE II
CORRELATION HEATMAP OF THE TODDLER DATASET

Results from the correlation heatmap confirms are doubts that some of the questions in the AQ-10 test do not conform to toddlers. Next, we explored the relationship between Age and ASD positiveness. As ASD is a developmental disorder Age was one of the most important factors. The Adult dataset included people ranging from ages 14 to 61, where are the Toddler dataset has age ranging from 0 to 36 months.

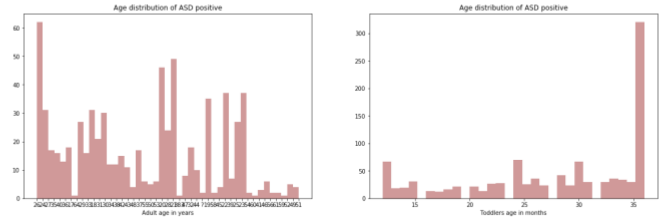

FIGURE III
AGE DISTRIBUTION OF ASD POSITIVE ADULTS AND TODDLERS

For Adults most of the ASD positive are around 20 or 30 years of age, whereas for toddlers most of them are around 36months. We can see in adults as the age increases the number decreases, whereas in toddlers as the age increases the number increases. It goes well with the research. For adults, people with autism develop strategies to help them age better. For toddlers, the significant signs of autism reveal around 3 years of age.

We additionally also found the relationship between ethnicity and Family member with ASD positiveness.

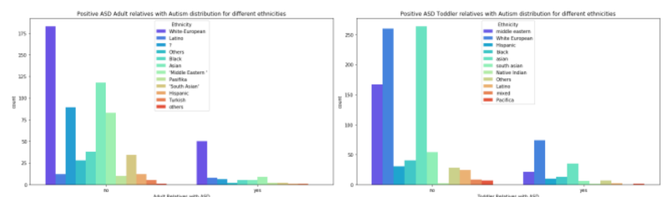

FIGURE IV
ASD POSITIVE RELATIVES WITH AUTISM FOR DIFFERENT ETHNICITIES

We can observe that both in Adults and Toddlers, White and Europeans Ethnicities have very high chance of being ASD positive if they have it in their genes. Black and Asians follow the next though with smaller ratios. We cannot conclude anything firmly, but we can say confidently that there is a genetic link for ASD positiveness as backed by studies.

For further data analysis data had to be label encoded into integers. We used Label Encoder to convert all the attributes to integers. The dataset is further divided into train and test dataset with a test size of 0.2 and a random state of 7.

We used several different machine learning algorithms like Linear Regression, Linear Discriminant Analysis, K-nearest Neighbor Classifier, Decision Tree Classifier, Gaussian Naïve Bayes Classifier, Support Vector Machine Classifier and Random Forest Classifier. We iterated all the above-mentioned algorithms on different parametric values to achieve the best possible model. We used GridSearchCV algorithm to achieve this.

Generating Machine Learning Models

I. Logistic Regression

With change in parameter 'C': [0.01,0.1,1,10,100,1000]. We ran over 60 models with different C parameters, fitting 10 folds for each of the 6 candidates. The best Logistic Regression Model with {'C' : 100} has an accuracy of 100%.

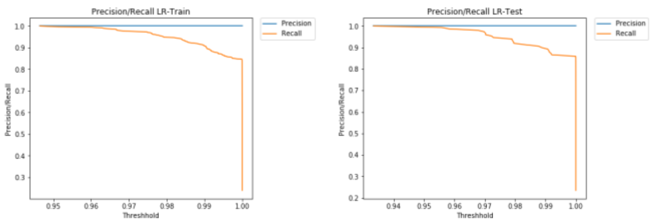

FIGURE V
PRECISION AND RECALL CURVES FOR TRAIN AND TEST DATASETS.

II. Linear Discriminant Analysis

With change in parameters 'solver': ['svd','lsqr','eigen']. We ran over 30 models with different C parameters, fitting 10 folds for each of the 3 candidates. The best Linear Discriminant Analysis with {'solver': 'svd'} has an accuracy of 95.2%

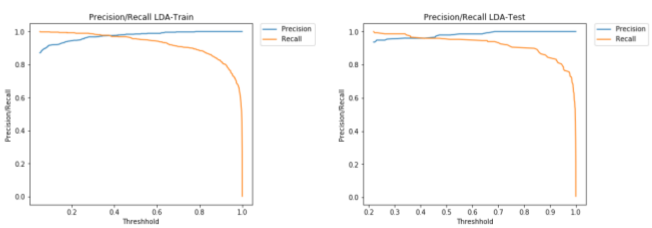

FIGURE VI
PRECISION AND RECALL CURVES FOR TRAIN AND TEST DATASETS.

III. K-Nearest Neighbors Classifier

With change in parameters 'n_neighbors': [1,10,50,100,500], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute']. We ran over 200 models with different C parameters, fitting 10 folds for each of the 20 candidates. The best K-nearest neighbor with {'algorithm': 'ball_tree', 'n_neighbors': 10} classifier has an accuracy 89.09%

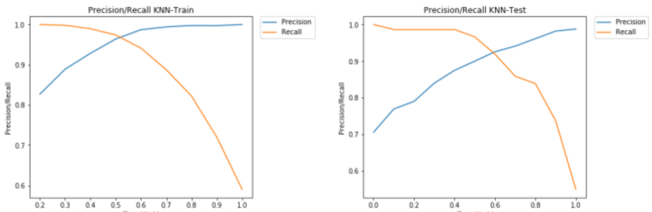

FIGURE VII
PRECISION AND RECALL CURVES FOR TRAIN AND TEST DATASETS.

IV. Decision Tree Classifier

With change in parameters 'criterion': ['gini', 'entropy'], 'min_samples_split': [2,3,4,5,6,7,8,9,10], 'min_samples_leaf': [1,2,3,4,5,6,7,8,9,10]. We ran over 1800 models with different C parameters, fitting 10 folds for each of the 180 candidates. The best Decision Tree Classifier with {'criterion': 'entropy', 'min_samples_leaf': 5, 'min_samples_split': 9} has an accuracy of 93.83%.

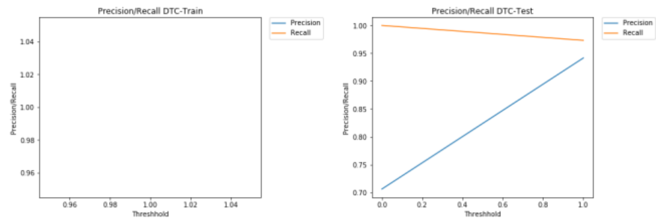

FIGURE VIII
PRECISION AND RECALL CURVES FOR TRAIN AND TEST DATASETS.

V. Gaussian Naïve Bayes Classifier

With change in parameters 'var-smoothing': [0.01,0.1,1,10,100,1000] we ran over 1800 models with different C parameters, fitting 10 folds for each of the 180 candidates. The best Gaussian Naïve Bayes classifier with {'C': 100} has an accuracy of 93.83%.

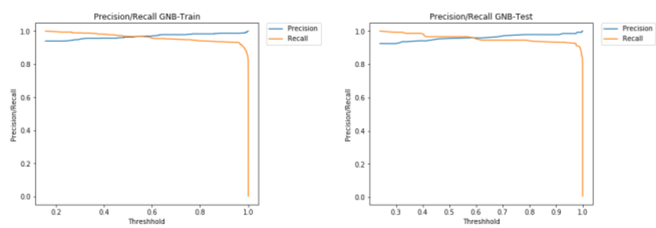

FIGURE IX
PRECISION AND RECALL CURVES FOR TRAIN AND TEST DATASETS.

VI. Support Vector Machine Classifier

With change in parameters 'C': [0.1,0.8,0.9,1,1.1,1.2,1.3,1.4], 'kernel': ['linear', 'rbf'], 'gamma': [0.1,0.8,0.9,1,1.1,1.2,1.3,1.4]
We ran over 1280 models with different C parameters, fitting 10 folds for each of the 128 candidates. The best Support Vector Machine classifier with {'C': 0.8, 'gamma': 0.1, 'kernel': 'linear'} classifier has an accuracy 100%

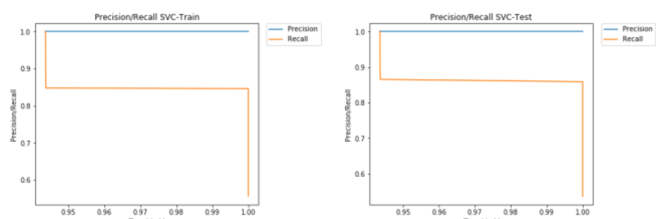

FIGURE X
PRECISION AND RECALL CURVES FOR TRAIN AND TEST DATASETS.

VII. Random Forest Classifier

With change in parameters 'n_estimators': [1,10,50,100,500], 'criterion': ['gini','entropy'], 'min_samples_split': [2,3,4,5,6,7,8,9,10], 'min_samples_leaf': [1,2,3,4,5,6,7,8,9,10]. We ran over 9000 models with different C parameters, fitting 10 folds for each of the 900 candidates. The best Random Forest classifier with {'criterion': 'entropy', 'min_samples_leaf': 6, 'min_samples_split': 9, 'n_estimators': 100} has an accuracy 100%.

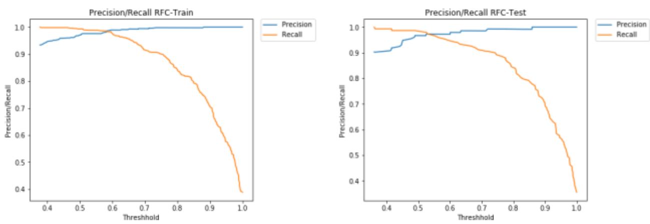

FIGURE XI
PRECISION AND RECALL CURVES FOR TRAIN AND TEST DATASETS.

PRECISION RECALL CURVE

There are many ways to evaluate the skill of a prediction model. An approach in the related field of information retrieval (finding documents based on queries) measures precision and recall. These measures are also useful in applied machine learning for evaluating binary classification models. Precision is a ratio of the number of true positives divided by the sum of the true positives and false positives. It describes how good a model is at predicting the positive class. Precision is referred to as the positive predictive value. The precision and recall can be calculated for thresholds using the precision_recall_curve() function that takes the true output values and the probabilities for the positive class as output and returns the precision, recall and threshold values. When comparing among models the model with the highest curve is considered to be a better performing model than that of a model with a lesser curve.

CONCLUSION

Looking at the Precision Recall Curve of the performing algorithms on test data, we can clearly say that Support Vector Machine Classifier and Random Forest Classifier are the best possible algorithms to predict ASD positiveness. And when Support Vector Machine Classifier and Random Forest Classifier are considered, it is Random Forest Classifier that has the better performance among both. Although Linear Regression Model did have a 100% accuracy, the algorithm is primitive and lacks the flexibility to handle complex multi-dimensional data.

Looking back at all the conclusions we reached and hypothesis we tested, it is clear to say that when a robust test is created as a viable replacement for AQ-10 test, using machine learning algorithms we can clearly predict as ASD positiveness and negativeness in both Adults and Toddlers. We also further suggest a series of behavioral test for Toddlers to confirm the positive ASD diagnosis and to find the extent of the disorder.

REFERENCES

[1] Thabtah, F. (2017). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.

[2] Thabtah, F. (2017). Machine Learning in Autistic Spectrum Disorder Behavioral Research: A Review. To Appear in Informatics for Health and Social Care Journal. December 2017

[3] Teates, J. (2018), Using Informatics to Improve Autism Screening in a Pediatric Primary Care Practice. DNP Scholarly Projects.

[4] Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th, 2017].

[5] Lord C, Risi S, DiLavore PS, Shulman C, Thurm A, Pickles A. (2006) External icon Autism from 2 to 9 years of age. Arch Gen Psychiatry. 2006 Jun;63(6):694-701.

[6] Gomotab, M., Bernard, F. A., Matthew, D. H., Belmonte, M. K., Ashwin, C., Bullmore, E. T., Baron-Cohen. S. (2006) Change detection in children with autism: An auditory event-related fMRI study. Elsevier.

[7] Johnson, C. P., Myers, S. M. (2007) Identification and Evaluation of Children with Autism Spectrum Disorders. the Council on Children with Disabilities. Pediatrics November 2007, 120 (5) 1183-1215; DOI: https://doi.org/10.1542/peds.2007-2361