International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

# Machine Learning Approach to the Classification Problem for Autism Spectrum Disorder

**Mandar Chougule[1], Hrushikesh Dambhare[2], Sheetal Chakral[3], Prof. Vandana Salve[4]**

[1, 2, 3] *Dept. Of Computer Engineering, M.G.M's College of Engineering and Technology, Kamothe, Navi Mumbai, Maharashtra, India*

[4] *Professor, Dept. Of Computer Engineering, M.G.M's College of Engineering and Technology, Kamothe, Navi Mumbai, Maharashtra, India*

---

**Abstract -** *Autistic Spectrum Disorder (ASD) is the name for a group of developmental disorders impacting the nervous system. Autistic Spectrum Disorder (ASD) is the name for a group of developmental disorders impacting the nervous system. ASD symptoms range from mild to severe: mainly language impairment, challenges in social interaction, and repetitive behaviors. Many other possible Symptoms include anxiety, mood disorders and Attention-Deficit/Hyperactivity Disorder (ADHD). ASD has a significant economic impact in the healthcare domain, both due to the increase in the number of ASD cases, and because of the time and costs involved in diagnosing a patient. Early detection of ASD can help both patients and the healthcare sector by prescribing patients the therapy and/or medication they need and thereby reducing the long-term costs associated with delayed diagnosis. However, challenges remain. Pursuing such research necessitates working with datasets that record information related to behavioral traits and other factors such as gender, age, ethnicity, etc. Such datasets are rare, making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. At present, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. These data are extremely sensitive and hard to collect for social and personal reasons and the regulations around them. Using Machine learning methods to find a predictive model which can be further improved to accurately provide the common populace a way to understand and identify those suffering from ASD.*

***Key Words***: ASD, Autism, ASD Screening Methods, ADHD, ASD Datasets, Machine Learning Models.

## 1. INTRODUCTION

I was able to find open-source data available at UCI Machine Learning Repository. The data was made available to the public on December 24th, 2017. The data set, which I will be referring to as the ASD data set from here on out, came with a .csv file that contains 704 instances that are described by 21 attributes, a mix of numerical and categorical variables. A short description of ASD dataset can be found on this page. This data set was denoted by Prof. Fadi Fayez Thabtah, Department of Digital Technology, MIT, Auckland, New Zealand, fadi.fayez@manukau.ac.nz.

### 1.1 Problem Statement

With the available ASD data on individuals my goal is to make predictions regarding new patients and classify them into one of two categories: "patient has ASD" or "patient does not have ASD".

In other words, we are working on a binary classification problem with the ultimate goal of being able to classify new instances, i.e., when we have a new adult patient with certain characteristics, we would like to able to predict whether or not that individual has high probability of having ASD.

This work aims to explore several competing supervised machine learning classification techniques namely:

1. Decision Trees
2. Random Forests
3. Support Vector Machines (SVM)
4. k-Nearest Neighbors (kNN)
5. Naive Bayes
6. Logistic Regression
7. Linear Discriminant Analysis (LDA)
8. Multi-Layer Perceptron (MLP)

### 1.2 Metrics

In order to choose the appropriate model that avoids Underfitting or Overfitting the data we will analyse the Bias-Variance Trade-Off, Model Complexity Graph, Learning Curves and Receiver Operator Characteristic Curves (ROC). To measure the effectiveness of each classification model we will study the accuracy score along with the precision, recall, F-Beta Score and confusion matrix.

Definition 1.1. Model Complexity Graph

A model complexity graph plots the training error and cross validation error as the model's complexity varies, i.e., the x-axis represents the complexity of the model (such as degree of the polynomial for regression or depth of a decision tree) and the y-axis measures the error in training and cross validation.

Figure 1 below shows a typical model complexity graph.

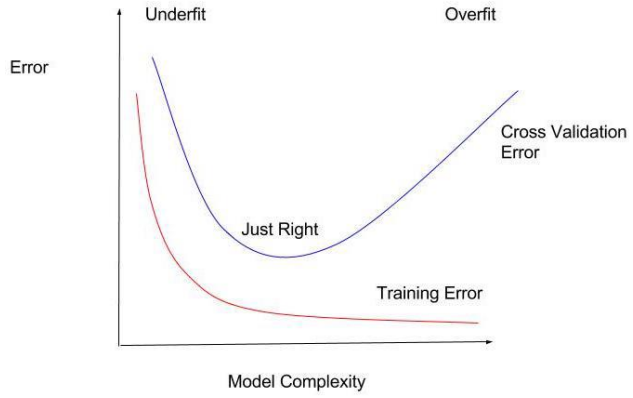


**Fig 1**: Model Complexity Graph

Definition 1.2. Accuracy

Accuracy measures how often the classifier makes the correct prediction. In other words, it is the ratio of the number of correct predictions to the total number of predictions (the number of test data points). Accuracy is needed as the fraction of correct predictions.
accuracy = true positive + true negative / true positive + false positive + false negative + true negative

Definition 1.3. Precision

Precision measures how accurate our positive predictions were i.e., out of all the points predicted to be positive how many of them were actually positive.
precision = true positive / true positive+false positive

Definition 1.4. Recall

Recall measures what fraction of the positives our model identified, i.e., out of the points that are labelled positive, how many of them were correctly predicted as positive. Another way to think about this is what fraction of the positive points were my model able to catch?
recall = true positive / true positive+false negative

## 2. ANALYSIS

After using machine learning algorithms to to clean the dataset and rid of any missing values, we can begin analyzing and exploring the dataset to efficiently create a machine learning model able to identify ASD patients.

### 2.1 Data Exploration

Our data set involves ten behavioral features (\AQ-10-Adult") (binary data) and ten individual characteristics such as \Gender", \Ethnicity", \Age", etc (categorical data) and one numerical data (\result"). Table 2) below lists all variables involved in the ASD data set.

**Table -1**: List of Attributes in ASD dataset.

| Variable Name | Description |
|---|---|
| Age | Age in years |
| Gender | male or female |
| Ethnicity | list of common ethnicities in text format |
| Born with Jaundice | whether case was born with jaundice |
| Family member with PDD | whether any immediate family member has a PDD |
| Who is completing the test | parent, self, caregiver, medical sta_, clinician, etc |
| Country of Residence | list of countries in text format |
| Used the screening app before | whether the user has used screening app |
| Screening Method Type | type of screening method chosen based on age category |
| Question 1 Answer | the answer code of the question based on the screening method used |
| Question 2 Answer | the answer code of the question based on the screening method used |
| Question 3 Answer | the answer code of the question based on the screening method used |
| Question 4 Answer | the answer code of the question based on the screening method used |
| Question 5 Answer | the answer code of the question based on the screening method used |
| Question 6 Answer | the answer code of the question based on the screening method used |
| Question 7 Answer | the answer code of the question based on the screening method used |
| Question 8 Answer | the answer code of the question based on the screening method used |
| Question 9 Answer | the answer code of the question based on the screening method used |
| Question 10 Answer | the answer code of the question based on the screening method used |
| Screening Score | _nal score obtained based on scoring algorithm of screening method used |

### 2.2 Algorithms and Techniques

### I. Support Vector Machines (SVM):

Support Vector Machine is a supervised machine learning algorithm that is commonly used in classification problems. It is based on the idea of finding the hyperplane that `best' splits a given data set into two classes. The algorithm gets its name from the support vectors, which are points of a data set that if removed would alter the position of the separating

hyperplane. which are points of a data set that if removed would alter the position of the separating hyperplane.

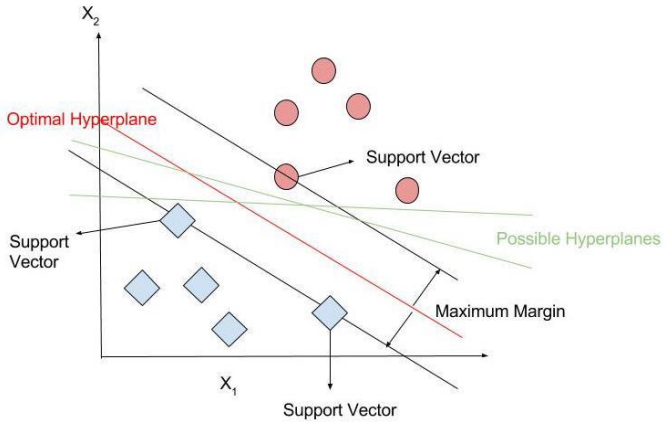


**Fig 2**: Support Vector Machine Diagram

## II. Logistic Regression:

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent variables.

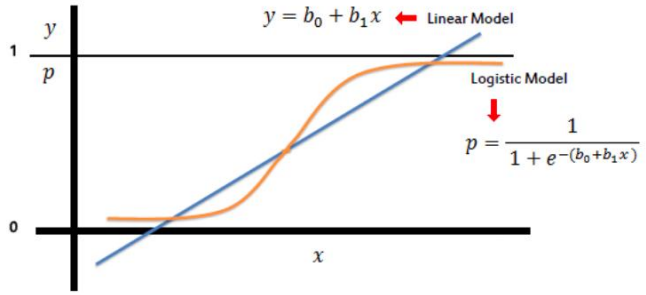


**Fig 3**: Logistic Regression Model

### 2.3 Benchmark Model

Our data set was released on the UCI repository on the 24th of December, 2017. Very little work has been done with this particular data. We are unaware of any rigorous data classification problems related to this ASD data using a machine learning approach but in there is some mention of a Decision Tree algorithm can be applied for this classification problem although no numerical results or measurable metric were presented in the article.

## 3. CONCLUSIONS

### 3.1 Free-Form Visualization

As in Feature Selection" exploration, we have seen the attribute named `result' has such a powerful presence in the ASD dataset, all other attributes have little to no contribution in deciding the final class. In the Figure below, we have drawn a swarmplot with `result' as x-axis and ASD-class (say `yes' is 1 and 'no' is 0) as y-axis, and reconfirmed the underlying association between the given variables where the target class is easily separable.

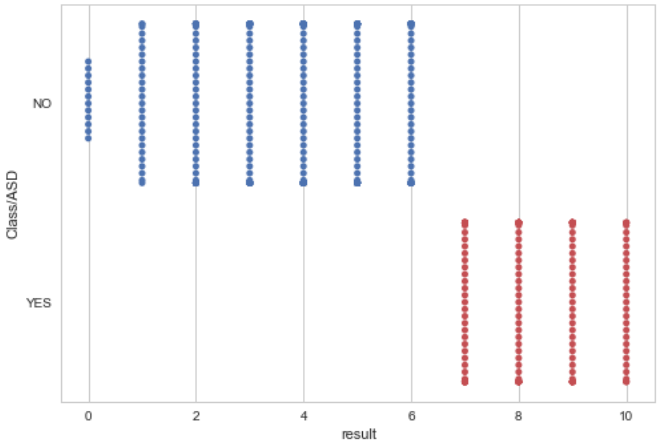


**Fig 4**: **Free-Form Visualization**

### 3.2 Reflection

During pre-processing of the data, we dropped about 95 of rows of data due to its `NaN' entries. Ideally one should try to retain as much data as possible, as there could always be some valuable insight that could be lost.

## 4. IMPROVEMENTS

In our consideration, to build an accurate and robust model, one needs to have larger datasets. Here the number of instances after cleaning the data were not sufficient enough to claim that this model is optimum. Looking at the performances of our learning models, nothing can be improved with this current data set as models are already at their best. After discussing this issue with a researcher directly working on adult autism, we have realized that it is extremely difficult to collect a lot of well documented data related to ASD.

This ASD dataset has recently been made public (available from December 2017), and thus not much work has been done. With this in consideration, our research has resulted in well-developed models that can accurately detect ASD in individuals with given attributes regarding the persons behavioural and medical information. These models can serve as benchmarks for any machine learning researcher/practitioner who is interested in exploring this dataset further or other data sets related to Autism screening disorder.

## REFERENCES

[1] Brian Godsey, Think Like a Data Scientist Manning, ISBN: 9781633430273

[2] H. Brink, J. Richards, M. Fetherolf, Real World Machine Learning, Manning, ISBN:9781617291920

[3] D. Cielen, A. Meysman, M. Ali, Introducing Data Science, Manning ISBN:9781633430037.

[4] J. Grus, Data Science From Scratch First Principles With Python, O'Reilly ISBN:9781491901427

[5] A. G_eron, Hands-On Machine Learning with Scikit-Learn & Tensor Flow, O'Reillyn.ISBN: 9781491962299

[6] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Second Edition, Springer