A FEATURE SELECTION BASED MACHINE LEARNING MODELS FOR THE DIAGNOSIS OF AUTISM SPECTRUM DISORDER

Nikita Jijo
Department of Mathematics
Vellore Institute of Technology
Vellore,India
nikita.jijo2021@vitstudent.ac.in

Sheethal Joseph
Department of Mathematics
Vellore Institute of Technology
Vellore,India
sheethal.joseph2021@vitstudent.ac.in

Rushi Kumar B
Department of Mathematics
Vellore Institute of Technology
Vellore,India
rushikumar@vit.ac.in

***Abstract*—Autism Spectrum is a developmental condition caused because of brain abnormalities. Recent studies indicate that 1 in 44 American children suffer from autism. Today's urgent need is for the advancement of more reliable diagnostic tools for early autism diagnosis. With the assistance of diverse machine learning approaches, this may be accomplished. The goal of this project is to examine the accuracy of implementing backward feature selection-based ML algorithms such as "SVM", "Random Forest", "Logistic Regression", "Naive Bayes", and "K- Nearest Neighbor" on three separate ASD datasets from the UCI Machine Learning Repository, each of which contains 21 features. We obtained that for all three data sets, Logistic Regression performs better both before and after the use of feature selection.**

Keywords: Autism spectrum disorder, Feature- Selection, Logistic Regression, Random Forest, Naïve Bayes, K-Nearest Neighbor(KNN)

I. INTRODUCTION

Autism Spectrum Disorder, often known as ASD, is a neural and generative state that makes noticeable impact on how persons behave, interact, and connect with others. Children with autism frequently exhibit symptoms before the age of three. Studies reveal that there is no one aetiology for autism, even though the root cause is still unclear. Autism may be brought on by a genetic disorder as well as environmental elements like having elderly parents or being underweight at birth. People with ASD might have a range of abilities. While some might not speak at all, others might have highly developed communication skills. While some people may need assistance from others on a daily basis, some people can work and live on their own. "ASD affects 31% of kids intellectually (intelligence quotient [IQ] <70), 25% are borderline ([IQ] 71-85), and 44% have IQs that are ordinary to above average ([IQ] >85)" [14]. Children with autism spectrum disorders may struggle with basic social interactions. Some signs and symptoms include:

- Engaging and connecting with others seems challenging
- Having trouble making eye contact
- Having difficulty in comprehending what other people are thinking or feeling
- Using a distinctive tone or rhythm when speaking
- Fails to responds to his or her name

The most crucial actions should be considered to lessen the indicators of ASD so as to enhance the wholesomeness for sufferers are primary detection and ministrations. Early screening is always advised by experts because it enables the early identification of children who are more likely to have autism spectrum disorder and encourages their families to take the necessary steps to get the children access to early intervention services. There is, however, no method or medical test for autism detection. Observation is usually how ASD symptoms are identified.

Autism may now be predicted at an early level because of the advances in artificial intelligence and machine learning. In addition to assisting in a quick and accurate assessment of ASD risk, machine learning techniques are needed for accelerating the whole diagnostic course and providing an helping hand for families to access the much required healings more quickly [16].

With the implementation of ML tools and the Autism Spectrum Quotient (AQ) as the screening tool, we aim to make a contribution to the early detection of ASD.

II. LITERATURE SURVEY

Suman Raja et al. [1] together published a paper which attempts to investigatethe necessity of various machine and deep learning tools like "Naive Bayes", "SVM", "Logistic Regression", "KNN", "Neural Network",and "Convolutional Neural Network" in an effort to anticipate and analyze ASD issues in 3 datasets. The conclusion interpreted that CNN-employed -models was most preferable on all datasets, with top accuracy of 99.53%, 98.30%, and 96.88%. These findings suggested that a CNN-based model rather than a conventional machine learning classifier could be employed to detect autism.

Vaishali R. et al. [2] have proposed to create machine learning-based behavioral analytics in order to detect the fear of autism quicker than outdated testing methods. For this, an ASD diagnosis dataset is tested using a binary firefly feature selection wrapper based on swarm intelligence. The trail's other hypothesis asserts that a ML model can pull off preferable classification correctness with not many feature subdivision to decide subset fitness. It was found that 10 of the 21 features in the ASD dataset are ample to pick out between who have ASD and people who do not have ASD using only a Swarm intelligence-based solitary binary firefly feature selection framework. The outcomes of their method verify the above-mentioned premise by generating an average accuracy of 92.12% to 97.95% with optimal feature subdivisions, which is almost identical to the average correctness brought out by the whole ASD identification dataset. The wrapper based on swarm intelligence, according to this article, is a preferable option to feature reduction techniques.

J.A Kosmicki et al. [3] conducted a investigation on a limited range of ways for autism identification using feature selection-employed machine learning. In this learning, they used machine learning to check out the Autism Diagnostic Observation Schedule (ADOS), the finest and broadly employed equipment for experimental examination of ASD, to determine if only a subgroup of behaviours can make a distinction between children having or not having the autism spectrum. ADOS made up of four units with segment 2 for people with little terminology and segment 3 for advanced levels of mental abilities. The outcomes support the fact that the fewer features when employed ML tools can have higher degree of accurateness in autism detection.

Ashima Sindhu Mohanty et al. [4] presented a innovative method for identifying early ASD using a deep classifier. In this study, an attempt is made to include Principal Component Analysis (PCA) for feature dimension trimming in the quantity of properties by the use of 10-fold cross validation, followed by the use of Deep Neural Network (DNN) for ASD class type classification. The different rating variables such as “sensitivity”, “accuracy”, “specificity” and “F-valuess” produced suitable results. The results of the experiment show that PCA in fusion with DNN provides clinically justifiable achievement for functional ASD recognition.

Basma Ramadan Gamal Elshoky et al. [5] investigated the numerous feature selection approaches on four ASD datasets for knowing important features for refining the ASD classification scheme. To rank significant features, several feature engineering techniques are used. The steps followed are filter, select features, dataset splitting, classification algorithms, time performance measurement, and performance. The correlation matrix routine demonstrated the relationship between features, allowing to take the most remarkable features. A number of machine learning classifiers are used. On different sizes of data, the picked- out features achieve 100% “correctness”, “specificity”, “sensitivity”, “AUC”, and “f1 score” using “adaboost”, “linear discriminant analysis”, and “logistic regression classifier”. Cross-validation with 10 k-fold was used to validate the results.

Md. Mokhlesur Rahman et al. [6] focuses on current works for classification and feature selection of ASD. They propose techniques for improving the execution speed of machine learning techniques when filtering complex data for conception and its application in the research field of ASD. This paper remarkably benefits the future works using a ML approach for categorization, dealing out the disproportionate data in autism. A ML algorithm will give propitious results in detecting ASD by lowering data dimensionality and selecting the suitable features of autism.

Md Delowar Hossain et al. [7] the target of this paper is to point out the dominant characteristics and analysing the diagnosis procedures for better diagnosis. They examined ASD datasets taken from toddlers, children, adolescents, and adults and tested cutting-edge classification and feature selection methods to find out the foremost classifier and feature set for the above-mentioned ASD datasets. Their observations gives the conclusion that the “multilayer perceptron (MLP)” classifier outrun every other measures and bring off the best accuracy. They added that the “relief F” feature selection technique worked the finest among every other and was also ranked as the best.

Kaushik Vakadkar et al. [8] aims to look whether a child can be detected with ASD at their early stages, which consequently helps in the diagnosis. They used different models to analyse the data and built the models for prediction based on the results. They have created an mechanized ASD forecasting model using the least behaviour sets contained in each dataset. “Logistic Regression” was found to be the most accurate of the five models tested on our dataset.

Kayleigh K. Hyde et al. [9] imparts a evaluation of 45 studies with applications of supervised machine learning in ASD, including classification and text analysis algorithms. The primary target is to check and express supervised ML trends in the ASD literature. In the 35 ASD reviewed research, “SVM” and “ADtree” were the extensively used algorithms, and in the 10 reviewed articles, “Naive Bayes”, "SVM", and “Random Forest” were the most popularly used.

Tania Akter1and et al. [10] collected early-disposed ASD datasets from “toddlers”, “children”, “adolescents”, and “adults” and put in several feature revision methods to these datasets, consisting “log”, “Z-score”, and “sine functions” (FT methods). The working was analyzed by the above datasets. On the above-mentioned datasets, “SVM”, “Adaboost”, “Glmboost” and “Adaboost” performed the best vital features that predicts ASD.

From all the related works mentioned it is clear that there is a clear demand to investigate the chance of using ML- models through feature selection for noticing ASD in community. The majority of the efforts debated employs traditional machine learning tactics, which limit their

interpretation. In this article, the effectiveness of multiple ML models before and feature selection were compared. Different models are developed and then compared for one and all data sets.

III. DATASET

Three different types of ASD datasets, each with 21 characteristics, were acquired from the "UCI machine learning repository". There are 292 cases in the first dataset that are connected to ASD screening in children [11], 292 instances in the second dataset that are connected to ASD screening in adults [13], and 104 cases in the third dataset that are connected to ASD screening in adolescents [12]. The screening procedure in the AQ test allots a point for each question. If the person receives a score higher than 6, it is determined that they may have an ASD trait and they are forwarded for more diagnostic testing. In the adult category in table 1, if response is "slightly agree" or "definitely agree" for question items 1, 7, 8, and 10, a point would been awarded; for the leftover questions, a point would be awarded for answers that are "slightly disagree" or "definitely disagree." For question numbers 1, 5, 8, and 10 in the adolescent group in table 2, if the response is "slightly agree" or "definitely agree," a point is awarded; for standing questions, a point is awarded for "slightly disagree" or "definitely disagree." For the child category in table 3, if the response is "slightly agree" or "definitely agree" to question items 1, 5, 7, and 10, then a point is awarded; for the leftover questions, a point is awarded to the response "slightly disagree" or "definitely disagree" [4] [17] [18]. The AQ surveys for the "adult", "adolescent", and "child" categories are shown in the tables below.

TABLE I.

No.	AQ questionnaires for adult.				
	AQ-10 Adult questionnaire	*Definitely agree*	*Slightly agree*	*Slightly disagree*	*Definitely disagree*
1	I frequently hear faint noises that others do not.				
2	Normally, I pay more attention to the big picture than the specifics.				
3	I have no trouble managing multiple tasks at once.				
4	I can swiftly resume what I was doing in the event of an interruption.				
5	I am good at "reading between the lines" when people are speaking to me.				
6	I know how to tell if someone listening to me is getting bored				
7	When I'm reading a story I find it difficult to work out the characters' intentions				
8	I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant etc)				
9	I find it easy to work out what someone is thinking or feeling just by looking at their face				
10	I find it difficult to work out people's intentions				

TABLE II.

No.	AQ questionnaires for adolescent.				
	AQ-10 Adolescents questionnaire	*Definitely agree*	*Slightly agree*	*Slightly disagree*	*Definitely disagree*
1	S/he notices patterns in things all the time				
2	Typically, s/he focuses more on the big picture than the specifics.				
3	In a social group, s/he can easily keep track of several different people's conversations				
4	If there is an interruption, s/he can switch back to what s/he was doing very quickly				
5	S/he frequently finds that s/he doesn't know how to keep a conversation going				
6	S/he is good at social chit-chat				
7	S/he used to enjoy pretending games with other kids when they were younger.				
8	S/he finds it difficult to imagine what it would be like to be someone else				
9	S/he finds social situations easy				
10	S/he finds it hard to make new friends				

TABLE III.

No.	AQ questionnaires for children.				
	AQ-10 Children questionnaire	*Definitely agree*	*Slightly agree*	*Slightly disagree*	*Definitely disagree*
1	S/he often notices small sounds when others do not				
2	Typically, s/he focuses more on the big picture than the specifics.				
3	S/he can readily follow the conversations of multiple persons in a social group.				
4	S/he has little trouble switching back and forth between several activities				
5	S/he struggles to maintain a discussion with classmates.				
6	S/he is good at social chit-chat				
7	When reading a story, the s/he has trouble figuring out the character's motivations or emotions.				
8	When s/he was in preschool, s/he used to enjoy playing games involving pretending with other children				
9	S/he finds it simple to determine someone's thoughts or feelings only by seeing their face.				
10	S/he finds it hard to make new friends				

The other common features are: age, gender, ethnicity,

whether been affected with jaundice, country of residence, family members with autism, Used the app before, result, age description, relation and class/ASD traits.

IV. METHODOLOGY

i. *Data Pre-Processing*

Data pre-processing is a data acquisition practice used to change the source data into a composition that is both sensible and functional. There may be a lot of useless information and missingness in the data. Large datasets must be prepared properly to allow for the interpretation of the data they contain. The datasets employed for this project have missing values that were handled using mean and mode imputation techniques. The dataset's categorical variables were transformed into numeric values with the use of python libraries. In order to standardize the characteristics in the data into a fixed range before applying machine learning methods, feature scaling was also carried out.

ii. *Machine Learning Algorithms*

With a ratio of 80:20, the dataset has been split up into training and testing sets. The accuracy of methods like "Naive Bayes", "Support Vector Machine", "Logistic Regression", "Random Forest", and "K-Nearest Neighbor" is then evaluated.

a. *Naïve Bayes*

A supervised learning procedure that is employed for classification problems which is rooted on "Bayes Theorem" are called as "Naïve Bayes". It is one among the efficient classification algorithms that facilitate the occurrence of brilliant predictive analytical models proficient in giving accurate conclusions.

b. *Support Vector Machine*

Collection of supervised learning measures for segregating data and carrying out regression analysis are called "Support vector machines". Its approach look for an N-dimensional space which distinctly categories the datum.

c. *Logistic Regression*

The procedure for customizing the likelihood of a distinct outcome when an input variable is given is said to be known as "Logistic Regression". When binary and linear classification is involved LR could be a better choice. [15].

d. *Random Forest*

A supervised learning approach which develops and merges several decision trees to form a "forest." is known as "Random Forest". It averages the solution to make improvement in forecasted accurateness of datasets.

e. *K-Nearest Neighbor*

To group a single data point, a supervised learning classifier which is non-parametric called "K-NN" is used. Using the mentioned algorithm, a data point is restricted after the sorting of all the current data. has been stored.

iii. *Feature Selection*

Features are the input variables that we provide to our machine learning models. Our dataset's columns each represent a feature. Making ensuring that we only use the necessary features will help us train an optimal model. When there are too many characteristics, the model may pick up on noise and uninteresting patterns. The process of selecting the crucial aspects of our data is known as feature selection. In order to compare the accuracy before and after feature selection, our technique includes training algorithms employing backwards feature elimination. Regression analysis use the technique of backwards elimination to pick a subdivision of explanatory attributes for the model. Backward elimination is applied on the model's primary and all explanatory attributes. The variable having greatest p-value is next taken out of the model.

V. RESULT AND INTERPRETATION

i. *Exploratory Data Analysis*

Fig.1 shows the percentage of three age groups (children, adolescents, and adults) who have autistic symptoms. The results of the ASD screening in children reveal that 48.3% of them have autism, the results in adolescents reveal that 36.7% do, and the results in adults reveal that 26.8% have autism.

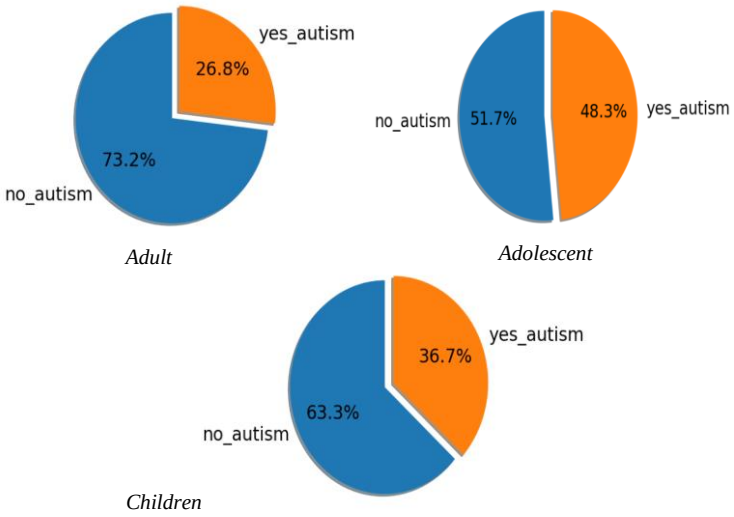

Fig.1. Percentage of types of age group having autistic symptoms

ii. *Model Evaluation*

The practise of employing different varieties of appraisal measures to figure out the ability of a predictive analytical model is said to be called as Model Evaluation. The common metrics for weighing classification achievement include

"accuracy", "precision", and "confusion matrix". The members of a confusion matrix are

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

The accuracy can be obtained by the formula-

$$\text{Accuracy} = \frac{TP+TN}{TN+TP+FN+FP}$$

For the ASD data of adults, adolescents, and children, the correctness of predictive analytics with and without feature selection has been examined. The three data sets' performance measures are displayed below.

TABLE I.

Results of ASD screening in children			
Algorithms	***Accuracy***		***Feature***
	Before	***After***	
Logistic Regression	100	100	10/21
Random Forest	100	100	9/21
Support Vector Machine	98.30	100	12/21
Naïve Bayes	94.92	96.61	17/21
K-Nearest Neighbor	81.36	83.05	15/21

The results of the asd screening for children are shown in table 4. Before implementing feature selection, various machine learning models on the dataset for diagnosing ASD children found accuracy ranging from (81.36% - 100%). KNN produced the lowest accuracy, 81.36%, whereas random forest and logistic regression both produced 100% accuracy. The accuracy attained after feature selection and training ML models with these picked features is between (83.05% - 100%).

TABLE II.

Results of ASD screening in Adolescents			
Algorithms	***Accuracy***		***Feature***
	Before	***After***	
Logistic Regression	100	95	13/21
Random Forest	85	80	16/21
Support Vector Machine	95	90	15/21
Naïve Bayes	80	85	16/21
K-Nearest Neighbor	80	80	16/21

The results of the asd screening for adolescents are shown in table 5. Before implementing feature selection, various machine learning models on the dataset for diagnosing ASD children found accuracy ranging from (80% - 100%). KNN and Naïve Bayes produced the lowest accuracy, 80%, whereas logistic regression produced 100% accuracy. The accuracy attained after feature selection and training ML models with these picked features is between (80%- 95%).

TABLE III.

Results of ASD screening in Adults			
Algorithms	***Accuracy***		***Feature***
	Before	***After***	
Logistic Regression	100	100	16/21
Random Forest	79	98.58	10/21
Support Vector Machine	100	97.87	9/21
Naïve Bayes	98.58	98.58	15/21
K-Nearest Neighbor	98.58	97.87	16/21

The results of the asd screening for adults are shown in table 6. Before implementing feature selection, various machine learning models on the dataset for diagnosing ASD children found accuracy ranging from (79% - 100%). Random Forest produced the lowest accuracy, 79%, whereas Logistic Regression produced 100% accuracy. The accuracy attained after feature selection and training ML models with these picked features is between (97.87% - 100%).

VI. CONCLUSION

Considerable machine learning approaches were used in this study to notice autism spectrum disorder.Using performance assessment criteria,several models were inked for ASD detection for 3 sets of data sets thst hsve been used . After applying feature selection to a dataset of children, Random Forest and SVM demonstrated 100% accuracy alongside Logistic Regression. Accuracy improvements were also seen with the other two approaches. But all other methods' accuracy decreased in adolescent ASD screenings, with the exception of Naive Bayes. After feature selection, adult ASD screening dataset accuracy increased. For all three data sets, Logistic Regression performs better both before and after the use of feature selection.

VII. REFERENCES

[1] S. Raj, S. Masood, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques", Procedia Computer Science, 167,994-1004, 2020.doi: 10.1016/j.procs.2020.03.399

[2] Vaishali R., and R. Sasikala, "A machine learning based approach to classify Autism with optimum behaviour sets", International Journal of Engineering & Technology 7(4): 18, 2018. doi: 10.14419/ijet.v7i3.18.14907Published

[3] J. A. Kosmicki, V. Sochat, M. Duda, and D. P. Wall, "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning," *Transl. Psychiatry,* vol. 5, no. 2, p. e514, 2015.

[4] A. S. Mohanty, P. Parida, K. C. Patra," Identification of Autism Spectrum Disorder using Deep Neural Network" *Journal of Physics: Conference Series,* Jan. 2006.

[5] B. R. Elshoky, O. A. Ibrahim, A. A. Ali, "Machine Learning Techniques Based on Feature Selection for Improving Autism Disease

Classification*", International Journal of Intelligent Computing and Information Sciences,* July 2021.

[6] Md. Rahman , O. L. Usman , R. C. Muniyandi,, S. Sahran, S. Mohamed and R. A. Razak ," A Review of Machine Learning Methods of Feature Selection and Classification for Autism Spectrum Disorder", *Brain Sciences*,10,949,2020.

[7] Md. Hossain, A. Anwar and Md. Zahidul Islam, "Detecting autism spectrum disorder using machine learning techniques", *Health Information Science and Systems* 9,17, 2021.

[8] Vakadkar K, Purkayastha D, Krishnan D, "Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques", *SN Computer Science*, 2(5): 386, 2021.

[9] K. K. Hyde, N. LaHaye, C. Parlett-Pelleriti, R. Anden, E. Linstead et al. "Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review". *Review Journal of Autism and Developmental Disorders* 6, 128–146, 2019.

[10] Tania Akter , Md. Shahriare Satu, Md. Imran Khan , Mohammad Hanif Ali , Shahadat Uddin , Pietro Lió , Julian M. W. Quinn, Mohammad Ali Moni, "Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorders", *in IEEE Access*, vol. 7, pp. 166509-166527, 2019.

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

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

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

[14] https://www.autism-society.org/what-is/facts-and-statistics/. Accessed 25 Dec 2019.

[15] T. W. Edgar, D. O. Manz, Chapter 4 - Exploratory Study, *Research Methods for Cyber Security*, pp. 95-130,2017.

[16] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi and M. N. Islam, "A Machine Learning Approach to Predict Autism Spectrum Disorder," *International Conference on Electrical, Computer and Communication Engineering (ECCE): IEEE*, pp. 1-6,2019.

[17] E. Weir, C. Allison, S. Baron-Cohen, "Identifying and managing autism in adults", vol. 31, pp.12-16,2020.

[18] E. Weir, C. Allison, S. Baron-Cohen, "Autism in children: improving screening, diagnosis and support", vol.31, pp.20-24,2020.