AUTISM PREDICTION USING MACHINE LEARNING

ABSTRACT

Autism Spectrum Disorder (ASD) has significant effects on language learning, speech, cognitive abilities, and social skills, often appearing in young children. Globally, ASD affects around 1% of the population, posing challenges for timely diagnosis and intervention due to reliance on clinical tests, leading to prolonged diagnostic processes and rising healthcare costs. To address these challenges, we investigate the application of machine learning algorithms like Support Vector Machines (SVM), Random Forest Classifier (RFC), Logistic Regression (LR), and Decision Tree (DT) to predict the likelihood of ASD development in early stages. By utilizing a comprehensive dataset, we build predictive models and compare their performance with traditional diagnostic methods. Our results demonstrate the effectiveness of machine learning in ASD prediction, with the Random Forest Classifier showing superior performance. This study underscores the potential of machine learning to streamline ASD diagnosis, potentially enabling earlier interventions and better outcomes for affected individuals.

INTRODUCTION

Autism Spectrum Disorder (ASD) presents a multifaceted challenge in the domain of neurodevelopmental disorders, significantly impacting an individual's communication, social interaction, and learning capabilities. While symptoms may emerge at any point in life, they typically become noticeable during early childhood and persist into adulthood. ASD encompasses a wide range of difficulties, including issues with focus, sensory sensitivities, and mental health concerns like anxiety and depression. The global prevalence of ASD has been on the rise, posing substantial barriers to timely diagnosis and intervention.

The World Health Organization (WHO) estimates that approximately 1 in 160 children worldwide is affected by ASD, underscoring the critical need for efficient screening and diagnostic methods. Diagnosing ASD is often a time-intensive and resource-demanding process, necessitating extensive assessments and evaluations. Yet, early identification of ASD is pivotal for initiating timely interventions and support, which can greatly enhance long-term outcomes for individuals with the disorder.

In response to these challenges, this project aims to develop a machine learning-based predictive model for ASD. The objective is to build an accurate and efficient tool capable of detecting autism traits across various age groups, thereby facilitating early screening and intervention efforts.

This project utilizes a comprehensive dataset and advanced machine learning algorithms to construct predictive models capable of identifying potential indicators of ASD. Through the development and evaluation of these models, we aim to contribute to the enhancement of ASD diagnosis and intervention strategies, ultimately leading to improved outcomes for individuals impacted by this condition.

LITERATURE REVIEW

This section reviews studies on ASD prediction techniques, emphasizing the efficacy of machine learning (ML) in diagnosing diseases based on symptomatology. For example, Cruz et al. [1] utilized ML for cancer and diabetes diagnosis, respectively. Wall et al. [2] employed an Alternating Decision Tree for

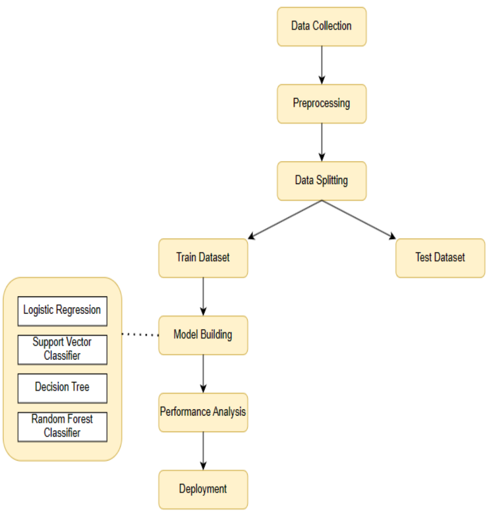
ASD trait detection with high accuracy using the ADI-R method, albeit limited to ages 5 to 17.

A few utilized SVM for ASD trait identification, achieving notable sensitivity but facing age range limitations (4-55 years). A few focused on screening tools and ML algorithm comparisons, respectively. Hauck and Kliewer highlighted significant screening questions for ASD diagnostic methods.

Bekerom [3] employed ML for ASD trait determination in children, while Wall et al. [4] emphasized the efficacy of short screening tests Others applied deep learning to identify ASD patients using brain imaging data. Liu explored ML algorithms for ASD identification using eye movement data, achieving promising results. Bone et al critically evaluated previous studies and reproduced results using their ML approach, identifying challenges in methodology and interpretation.

Despite extensive research, there's no consensus on a universally applicable ML-based autism screening tool across age groups, highlighting the need for diverse solutions, including app-based approaches.

FLOW DIAGRAM



METHODOLOGY

A. DATA COLLECTION

To construct a robust predictive model, we utilized the AQ-10 dataset, comprising three distinct subsets based on the AQ-10 screening tool questions. These subsets correspond to age groups: child, adolescent, adult. The AQ-10 tool is employed to determine if an individual warrants further assessment for autism spectrum disorder (ASD). Its screening questions cover various domains such as attention to detail, attention switching, communication, imagination, and social interaction. Each question is scored on a scale of 0 to 1, with 1 point assigned for affirmative responses. The dataset has one thousand instances and comprises twenty-two attributes, encompassing both numerical and categorical data, including age, gender, ethnicity, history of jaundice at birth, country of residence, prior use of the screening app, screening method type, individual questionnaire responses, results, and class labels.

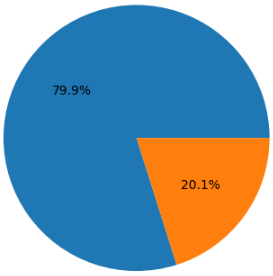
B. DATA SYNTHESIZATION

The data underwent synthesis to enhance its quality by removing irrelevant features. For instance, the column regarding previous app usage ("used_app_before") was excluded as it was deemed non-contributory to prediction model development. Moreover, categorical values underwent standardization to ensure consistency and facilitate analysis. Responses like "yes" and "no" were converted to numerical equivalents (1 and 0, respectively) for quantitative assessment. Ambiguous or undefined responses such as "?" or "others" were replaced with a standardized category label, such as "Others," for clarity and uniformity.

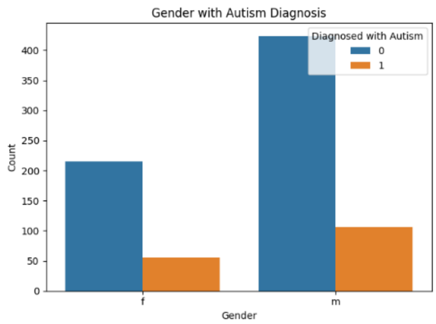
This synthesis streamlined the dataset, ensuring that only pertinent information remained for accurate prediction of autism spectrum disorder traits.

C. DATA VISUALISATION

The following chart depicts blue color for the people with no ASD traits in the training dataset and the orange color depicts the people with observed ASD traits in the training dataset which has eight hundred instances.



The below graph shows males have more ASD traits than females.



This way we can plot the graphs for all the other features and get to know the distribution and how the results get affected.

D. DEVELOPING THE PREDICTION MODEL

Begin by importing the necessary packages and reading the dataset into a Data Frame. The train-test split method is used to evaluate machine learning algorithms' performance in predicting autism traits. This method allows for a quick and straightforward assessment of different techniques' effectiveness in handling predictive tasks.

Convert string columns into binary format using a label encoder, while numerical columns are transformed using a one-hot encoder. Split the dataset into two groups: one containing all features except the class label, and the other containing only the class label.

1. Decision Tree: DT recursively partitions data based on significant attributes, forming a tree structure where each node represents a decision and each leaf node represents a class label. While interpretable, DTs are prone to overfitting.

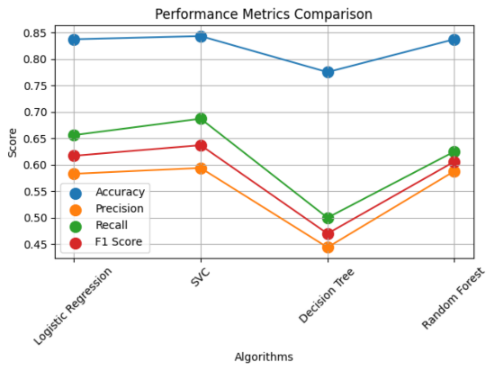
2. Random Forest Classifier: RF constructs multiple decision trees and outputs the mode of classes or the mean prediction. By introducing randomness in the tree-building process, RF mitigates overfitting and achieves higher accuracy.

3. Support Vector Classifier: SVC aims to find the hyperplane that best separates classes in feature space, maximizing the margin between classes. It works well in high-dimensional spaces and offers flexibility with different kernel functions for non-linear decision boundaries.

4. Logistic Regression: Despite its linear nature, logistic regression predicts class probabilities using a logistic function. It's computationally efficient, interpretable, and suitable for linearly separable data.

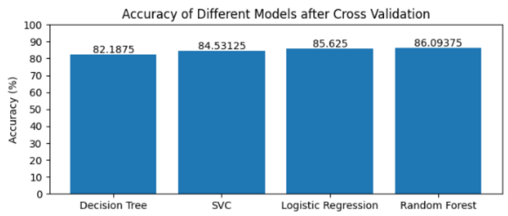
Random Forest emerged as the most promising algorithm in this evaluation, outperforming Logistic Regression, Support Vector Classifier, and Decision Tree methods.

Metrics	Logistic Regression	Support Vector Classifier	Decision Tree	Random Forest
Accuracy	0.837	0.843	0.775	0.837
Precision	0.583	0.594	0.444	0.588
Recall	0.656	0.687	0.500	0.625
F1 score	0.617	0.637	0.470	0.606



E. EVALUATING THE MODEL

Algorithms had been developed, and their accuracy had been evaluated, to generate predictions of autism traits. The models were evaluated using 10-fold cross-validation to assess their generalization performance. The Random Forest model consistently outperformed other models across all performance metrics.



CONCLUSION

ASD assessment is time-consuming due to overlapping symptoms, lack of quick diagnostic tools. We built an automated ASD prediction model using minimal behavior sets from diagnostic datasets, finding Logistic Regression most accurate among tested models. Limited large ASD datasets hindered accuracy. Nonetheless, our study offers insights for automated tools aiding childhood autism detection. Future plans include leveraging larger datasets and exploring deep learning techniques like CNNs for improved performance. Overall, our analysis lays groundwork for accurate ASD detection in children, serving as a resource for further research on ASD datasets.

REFERENCES

[1] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer informatics, vol. 2, 2006

[2] D. P. Wall, R. Dally, R. Luyster, J.-Y. Jung, and T. F. DeLuca, "Use of artificial intelligence to shorten the behavioral diagnosis of autism," PloS one, vol. 7, no. 8, p. e43855, 2012

[3] B. van den Bekerom, "Using machine learning for detection of autism spectrum disorder," 2017.

[4] D. Wall, J. Kosmicki, T. Deluca, E. Harstad, and V. Fusaro, "Use of machine learning to shorten observation-based screening and diagnosis of autism," Translational Psychiatry, vol. 2, no. 4, p. e100, 2012

[5]https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8679454&tag=1

[6]https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10040300