Project Report – Prediction of Autism Spectrum Disorder using Machine Learning

Abstract:

Autism Spectrum Disorder (ASD) profoundly impacts language learning, speech, cognitive, and social skills, often emerging in early childhood. With a global prevalence of approximately 1%, ASD presents challenges for diagnosis and intervention. Current diagnostic methods rely on clinical standardized tests, leading to prolonged diagnostic times and escalating medical costs. To address these issues, machine learning techniques offer promising avenues. In this study, we explore the use of machine learning algorithms such as Support Vector Machines (SVM), Random Forest Classifier (RFC), Logistic Regression (LR), and Decision Tree (DT) to predict the likelihood of individuals developing ASD in early developmental stages. Leveraging a comprehensive dataset, we develop predictive models and evaluate their performance against conventional diagnostic methods. Our findings demonstrate the effectiveness of machine learning models in ASD prediction, with the Random Forest Classifier emerging as the top performer. This study highlights the potential of machine learning to expedite ASD diagnosis, offering the possibility of earlier intervention and improved outcomes for affected individuals.

Introduction:

Autism Spectrum Disorder (ASD) presents a complex challenge within the realm of neurodevelopmental disorders, profoundly affecting an individual's communication, social interaction, and learning abilities. While symptoms can manifest at any age, they typically become evident in early childhood and persist throughout life. ASD encompasses a broad spectrum of challenges, including difficulties with concentration, sensory sensitivities, and mental health issues such as anxiety and depression. The global prevalence of ASD has been steadily increasing, posing significant obstacles for diagnosis and intervention.

According to the World Health Organization (WHO), approximately 1 in 160 children worldwide is affected by ASD, highlighting the urgent need for efficient screening and diagnostic tools. Diagnosing ASD is often a time-consuming and resource-intensive process, requiring extensive assessments and evaluations. However, early detection of ASD is crucial for initiating timely interventions and support, which can substantially improve 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 goal is to create an accurate and efficient tool capable of identifying autism traits across different age groups, facilitating early screening and intervention.

This project leverages a comprehensive dataset and advanced machine learning algorithms to construct predictive models capable of identifying potential indicators of ASD. By developing and evaluating these models, we seek to contribute to the advancement of ASD diagnosis and intervention strategies, ultimately improving outcomes for individuals affected by this condition.

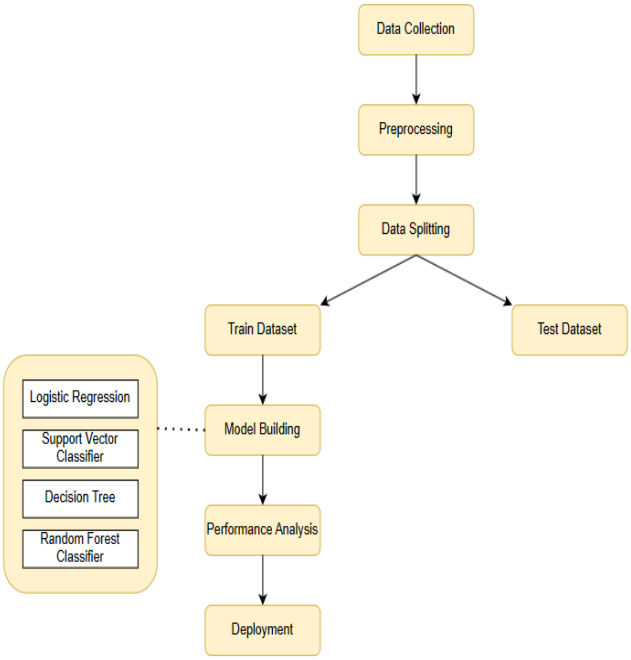
The subsequent sections of this project report will delve into the methodology employed, the implementation of the predictive model, and the evaluation of its performance. Through this endeavor, we aim to address the pressing need for efficient ASD screening tools and contribute to enhancing the quality of life for individuals affected by this condition.

Existing Methods:

Several studies have made use of machine learning in various ways to improve and speed up the diagnosis of ASD. A few applied forward feature selections coupled with under-sampling to differentiate between autism and ADHD with the help of a Social Responsiveness Scale containing 65 items. A few used metrics based on brain activity to predict ASD. Soft computing techniques such as probabilistic reasoning, artificial neural networks (ANN), and classifier combination have also been used. Many of the studies performed have talked of automated ML models that only depend on characteristics as input features. A few studies relied on data from brain neuroimaging as well. In the ABIDE database, few have extracted 6 personal characteristics from 851 subjects and performed the implementation of a cross-validation strategy for the training and testing of the ML models. This was used to classify patients with and without ASD, respectively. Others proposed a new ML technique called Rules-Machine Learning (RML) that offers users a knowledge base of rules for understanding the underlying reasons behind the classification, in addition to detecting ASD traits. Al Banna MH a researcher made use of a personalized AI-based system that assists with the monitoring and support of ASD patients, helping them cope with the COVID-19 pandemic.

In this study, we have used four ML models to classify individual subjects as having ASD or No-ASD, by making use of various features, such as age, sex, ethnicity, etc., and evaluated each classifier to determine the best-performing model.

Block Diagram:



About the Dataset:

To build our predictive model, we relied on the AQ-10 dataset, which is divided into three age groups: children, adolescents, and adults. This dataset evaluates various aspects related to autism using 10 screening questions. Each question allows respondents to score 0 or 1 point based on their answers. With 1000 instances and 22 features, including 8 categorical attributes, preprocessing was necessary to enhance data quality and model performance. The Dataset was later split into training 800 and testing 200 in an 80%-20% ratio.

Data Preprocessing:

In the data preprocessing stage, we began by removing irrelevant columns from the dataset, ensuring that only features relevant to the predictive task remained. Additionally, to handle missing values, we replaced any instances of "?" in the "ethnicity" column with the category "Others" to maintain data integrity.

Following this, since our dataset contained categorical variables, we converted them into numerical format using one-hot encoding. This transformation ensures that the categorical variables are represented in a format suitable for machine learning algorithms, allowing them to process the data effectively.

It's worth noting that our dataset had no missing values, which streamlined the preprocessing process and ensured that we could proceed directly to feature encoding without the need for imputation or other missing data-handling techniques. This clean and structured dataset serves as a solid foundation for building our predictive model for Autism Spectrum Disorder.

Classification Algorithms:

In our classification task, we employed four distinct algorithms.

1. Decision Tree (DT): DT recursively partitions data into subsets based on significant attributes at each node. It constructs a tree-like structure where each node represents a "decision" based on an attribute, and each leaf node represents a class label. While prone to overfitting, DTs are easy to interpret.

2. Random Forest Classifier (RF): RF is an ensemble method constructing multiple decision trees during training and outputs the mode of classes or mean prediction of trees. It introduces randomness in the tree-building process, using subsets of features and data samples. RF mitigates overfitting and tends to yield higher accuracy.

3. Support Vector Classifier (SVC): SVC finds the hyperplane that best separates classes in feature space, aiming to maximize the margin between classes for enhanced generalization. Effective in high-dimensional spaces, SVC is versatile due to its ability to use different kernel functions for non-linear decision boundaries.

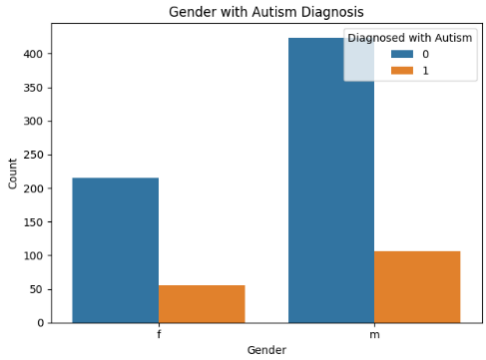
4. Logistic Regression: Despite its name, logistic regression is a linear model predicting the probability of a class. It uses the logistic function to map the output of a linear combination of features to a probability score. Logistic regression is computationally efficient, interpretable, and works well for linearly separable data.

Leveraging a combination of these classifiers, we aimed to develop a robust and accurate predictive model for autism spectrum disorder detection.

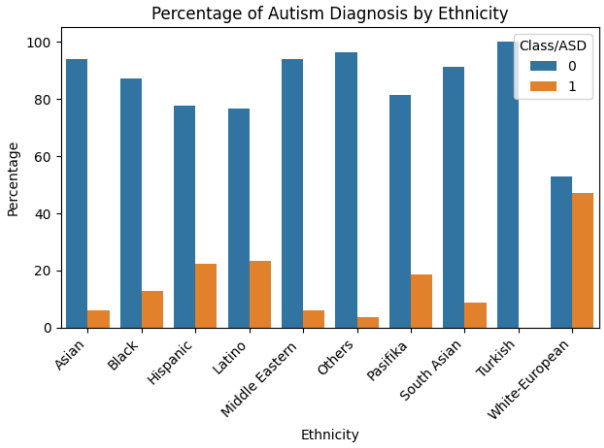
Data Analysis:

The dataset used here is based on the Quantitative Checklist for Autism in Toddlers (Q-CHAT) screening method. A shortened version, Q-CHAT-10, containing a set of 10 questions has been used. The answers to these questions are mapped to binary values as class type. These values are assigned during the data collection process using answering the Q-CHAT-10 questionnaire. The class value "Yes" is assigned if the score happens to be greater than 3, that is, there are potential ASD traits. Otherwise, the class value "No" is assigned, implying no ASD traits.

We plotted several graphs to get different visual perspectives of the dataset. It can be concluded that ASD is more prevalent in males than in females by observing the graph below.



The ethnicity distribution graph reveals that all Turkish individuals have no ASD traits. Among the given countries white-Europeans have the highest around (45%) observed ASD traits followed by Latinos and Hispanics.

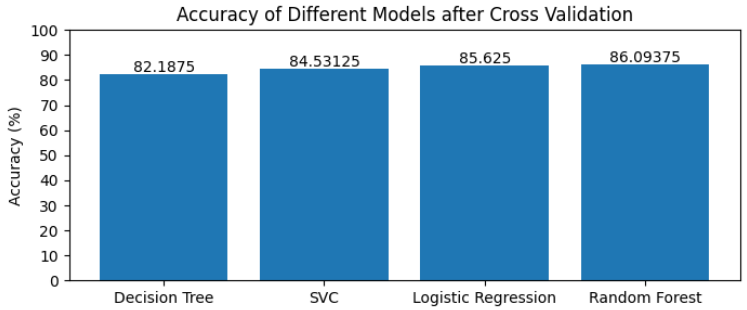


Results with discussion:

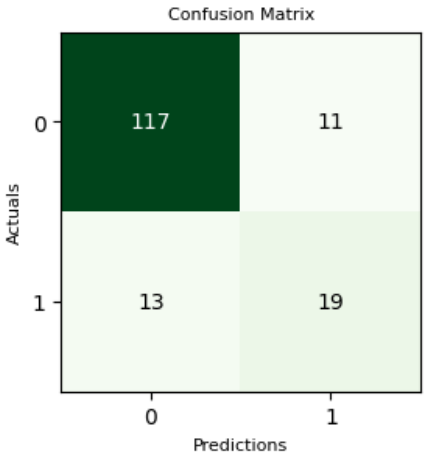
The four models while rigorously trained and the results are as follows. The Random Forest model demonstrated superior performance compared to other models, achieving the highest average accuracy, precision, recall, and F1 score. This indicates its robustness and effectiveness in predicting autism diagnosis. The Random Forest model stood out as the best-performing model due to its ensemble nature, which enables it to mitigate overfitting and improve prediction accuracy.

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

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.



Evaluation Matrix for Random Forest Classifier(best model) :



Interpretation:

True Positives (TP): The model correctly predicted 20 instances of autism cases. These are individuals who have autism, and the model accurately classified them as such.

True Negatives (TN): The model correctly predicted 114 instances of non-autism cases. These are individuals who do not have autism, and the model accurately classified them as such.

False Positives (FP): There were 14 instances where the model incorrectly predicted non-autism cases as autism. These false positives represent cases where individuals were incorrectly classified as having autism when they do not.

False Negatives (FN): The model incorrectly predicted 12 instances of autism cases as non-autism. These false negatives represent cases where individuals with autism were incorrectly classified as not having the condition.

The above four categories when put together in the form of a matrix produce the confusion matrix. The confusion matrix is particularly useful in gauging the performance of a machine-learning classification model. The Random Forest model demonstrates good performance in correctly identifying individuals without autism (TN) and those with autism (TP). However, the presence of false positives (FP) and false negatives (FN) indicates areas where the model can be further optimized. False positives (FP) could lead to unnecessary concern or interventions for individuals incorrectly classified as having autism. False negatives (FN) may result in missed opportunities for early intervention or support for individuals with autism who are not correctly identified by the model.

References :

[1] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9767927&tag=1

[2] https://link.springer.com/article/10.1007/s42979-021-00776-5