IMARTICUS
LEARNING

Post Graduation in
Data Science andAnalytics

Autism Spectrum Disorder in Toddlers

supervised by: Prof. Rithik Raj Vaishya
submitted on: 04-01-2024
submitted by: Prerak Jain

Contents

List of Tables 3

Abstract 4

1. Introduction 5

    1.1 Objective 8

    1.2 Motivation 9

    1.3 Background 11

2. Work Description 13

3. Technical Specification 15

4. Literature Survey 20

5. Methodology 22

6. Output 27

7. Conclusion 31

8. Reference 32

List of Figures

Figure 1:Autism prevalence estimates (Source: CDC reports [1])........ 5
Figure 2: ASD and its features........ 7
Figure 3: Common symptoms of ASD........ 11
Figure 4: Evaluation metrics formulas........ 19
Figure 5: Confusion Matrix........ 19
Figure 6: Machine learning life cycle model........ 22
Figure 7:Sigmoid function........ 25
Figure 8: AUC-ROC curve........ 30

List of Tables

Table 1: Dataset Description (source: [14])........ 28
Table 2: Algorithm and Accuracies........ 29
Table 3:Logistic Regression Result........ 29

Abstract

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by challenges in social communication, repetitive behaviors, and restricted interests. Early diagnosis and intervention are crucial for improving outcomes, yet ASD diagnosis often occurs later than desirable. In recent years, machine learning (ML) techniques have shown promise in aiding the early detection and characterization of ASD, offering a data-driven approach to complement traditional diagnostic methods. This report provides an overview of the current landscape of machine learning applications in ASD research. It discusses the use of ML algorithms for analyzing diverse data sources, including behavioral assessments, neuroimaging, and genetic data, to identify patterns and features associated with ASD. In conclusion, machine learning holds significant promise in advancing our understanding of ASD and improving early detection methods. As technology continues to evolve, integrating machine learning approaches into clinical practice may enhance the accuracy and efficiency of ASD diagnosis, ultimately leading to more timely interventions and improved outcomes for individuals on the autism spectrum.

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition that impacts the way a person perceives others and socializes with them. It affects three main developmental areas which are communication, social interaction, and repetitive patterns of behavior. As shown in Figure 1, Autism rates continue to rise dramatically, and according to Center for Disease Control (CDC) reports [1], prevalence rate increases with 1 in 54 children diagnosed with autism in 2020 compared to 1 in 59 in 2018, 1 in 110 in 2006, and 1 in 150 in 2000.

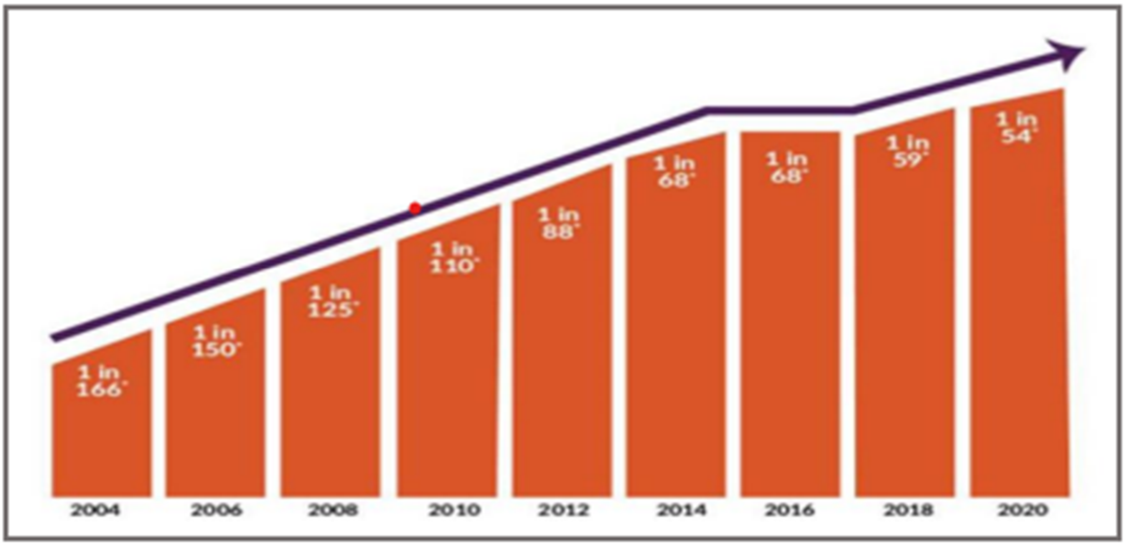

Figure 1:Autism prevalence estimates (Source: CDC reports [1])

Research has shown that there is no cure for ASD, however, early diagnosis and intensive intervention can make a big difference in the lives of many children and their families. Early diagnosis is very important for children on the spectrum because it allows to teach them the skills and behaviors that they lack at an early age when they still have good brain plasticity and therefore, the impact of intervention can be maximized and helps them reach their full potential. Interventions such as special education, behavior modification techniques, speech and occupational therapies help to bridge the gap that ASD kids have compared to their peers and speed up their development [2]. Therefore, early identification and diagnosis of children with ASD is very beneficial for the kids and their families and early screening is always recommended by experts because it allows to detect at an early stage the kids with more risk to be on the autism spectrum disorder, and thus, pushes their families to take the required measures to get the kids access the early intervention services.

Early screening of ASD has an important role in improving prognosis via early diagnosis and intervention [3]. For these kids with autism, both research and practice have shown that there is no magic pill for cure and early intervention through intensive education and behavior modification is the key due to its capability of changing the quality of life of these kids and their families and improving long-term outcomes [4]. However, the standardized tests for diagnosis such as Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview Revised (ADI-R), among others are time consuming, very costly and can be run only by trained clinicians. Thus, there are many barriers of getting these kids screened for autism at an early stage which contribute to delaying the therapies and interventions that they require. Nowadays, due to the advancement made in artificial intelligence and machine learning, autism can be predicted at an incredibly early stage [5]. Having access to an accurate, automated, cost effective and fast instrument for ASD screening at an early age will be greatly beneficial for detecting autism traits in the kids and securing a much better future for them.

Understanding the etiology of ASD remains a complex challenge. While a genetic predisposition is recognized, environmental factors and gene-environment interactions also play a role. The heterogeneity of ASD poses a challenge in pinpointing specific causes and effective interventions. Consequently, the management of ASD involves a multidisciplinary approach, incorporating behavioral therapies, educational interventions, and sometimes pharmacological treatments.

This project aims to explore the intersection of autism spectrum disorder and machine learning, seeking to leverage the power of computational methods to improve diagnostic accuracy and contribute to a deeper understanding of the heterogeneity within the ASD population. By harnessing the potential of machine learning, this research seeks to address the limitations of traditional diagnostic methods and pave the way for more efficient, objective, and personalized interventions for individuals on the autism spectrum.

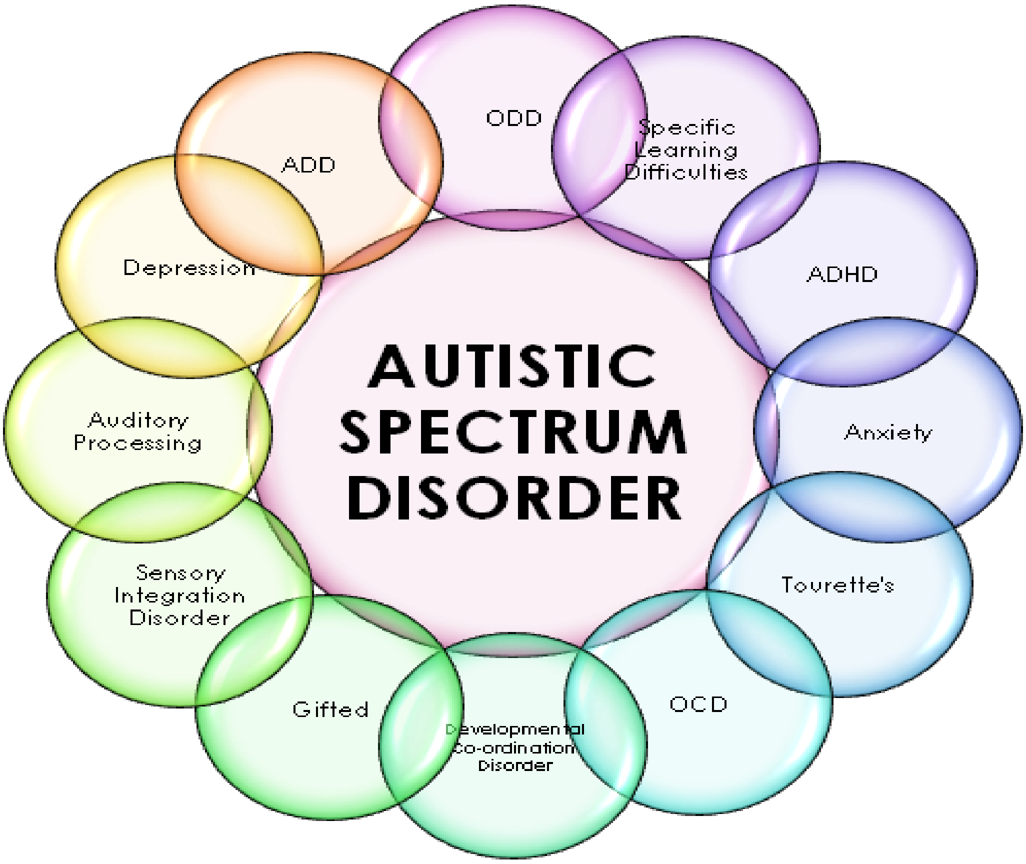
AUTISTIC SPECTRUM DISORDER
ODD
Specific Learning Difficulties
ADHD
Anxiety
Tourette's
OCD
Developmental Co-ordination Disorder
Gifted
Sensory Integration Disorder
Auditory Processing
Depression
ADD

Figure 2: ASD and its features

In the subsequent sections of this report, we will delve into the current landscape of ASD diagnosis, the challenges associated with traditional methods, and the potential advantages offered by machine learning. The project will also explore specific machine learning applications in ASD research, emphasizing the diverse data sources and modalities used to develop predictive models. Additionally, ethical considerations, challenges, and future directions in the integration of machine learning in ASD research and diagnosis will be discussed. Through this interdisciplinary approach, the project aspires to contribute to the ongoing efforts aimed at improving the lives of individuals with ASD and their families.

1.1 Objective

The primary objective of this project is to leverage machine learning techniques for enhancing the early detection, characterization, and understanding of Autism Spectrum Disorder (ASD). The specific goals of the project are as follows:

1. **Development of Predictive Models:**
Utilize machine learning algorithms to develop predictive models for early detection of ASD. Explore diverse data sources, including behavioral assessments, neuroimaging data, and genetic information, to identify patterns and features associated with ASD.
2. **Improvement of Diagnostic Accuracy:**
Evaluate the effectiveness of machine learning models in improving the accuracy of ASD diagnosis compared to traditional methods. Assess the potential of machine learning to identify subtle behavioral cues and patterns that may not be easily observable through conventional diagnostic approaches.
3. **Multimodal Data Integration:**
Investigate the integration of multiple modalities, such as behavioral, neuroimaging, and genetic data, to create comprehensive models that capture the heterogeneity of ASD. Explore how combining different types of data can contribute to a more nuanced understanding of the disorder.
4. **Ethical Considerations and Bias Mitigation:**
Address ethical considerations associated with the use of machine learning in ASD research, including privacy concerns, data security, and potential biases in algorithmic predictions. Implement strategies to mitigate biases and ensure responsible and ethical use of machine learning in the context of ASD diagnosis.
5. **Interdisciplinary Collaboration:**
Foster collaboration between researchers, clinicians, and data scientists to integrate machine learning approaches into the field of ASD research and diagnosis. Promote knowledge exchange and communication to bridge the gap between clinical expertise and technological advancements.
6. **Validation of Machine Learning Models:**
Validate the developed machine learning models on independent datasets to assess their generalizability and robustness. Evaluate the models' performance across diverse populations, age groups, and cultural contexts to ensure their applicability in a real-world clinical setting.
7. **Contribution to Biological Understanding:**
Investigate the potential of machine learning to contribute to our understanding of the biological underpinnings of ASD. Identify biomarkers and patterns in neuroimaging and genetic data that may shed light on the neurobiological mechanisms associated with ASD.

8. **Dissemination of Findings:**
Disseminate the research findings through academic publications, conferences, and other relevant channels. Share insights gained from the project with the scientific community to contribute to the ongoing discourse on the intersection of ASD and machine learning.

By achieving these objectives, this project aims to advance the field of ASD research, improve diagnostic practices, and provide valuable insights that can inform personalized interventions and support for individuals on the autism spectrum.

1.2 Motivation

The motivation stems from a deep recognition of the challenges faced by those on the autism spectrum, ranging from difficulties in social communication to unique sensory experiences. ASD's pervasive nature emphasizes the need for research and interventions that not only enhance diagnostic accuracy but also contribute to the development of tailored support strategies. The desire to improve the quality of life for individuals with ASD is a driving force, fueled by the potential to uncover the intricacies of this neurodevelopmental condition and foster a more inclusive and supportive society. The motivation behind undertaking a project focused on the intersection of Autism Spectrum Disorder (ASD) and machine learning is driven by several critical factors:

1. **Early Intervention Impact:**
Early detection and intervention significantly impact the developmental trajectory of individuals with ASD. Machine learning has the potential to enhance early detection by analyzing subtle patterns and behaviors, enabling timely and targeted interventions that can improve outcomes.

2. **Diagnostic Challenges with Traditional Methods:**
Traditional ASD diagnosis relies heavily on subjective clinical observations, leading to delays and potential misdiagnoses. Machine learning offers an opportunity to overcome these limitations by providing a data-driven, objective approach that can complement and enhance the accuracy of traditional diagnostic methods.

3. **Heterogeneity of ASD:**
ASD is characterized by its heterogeneity, with individuals presenting a wide range of symptoms and challenges. Machine learning allows for the integration of diverse data sources, such as behavioral, neuroimaging, and genetic data, enabling a more comprehensive understanding of the varied manifestations of

ASD.

4. **Potential for Personalized Interventions:**
Machine learning can contribute to the development of personalized interventions by identifying specific patterns and characteristics associated with individual cases of ASD. This has the potential to tailor interventions to the unique needs of each individual, optimizing the effectiveness of therapeutic strategies.

5. **Advancements in Technology:**
Rapid advancements in technology, particularly in machine learning and artificial intelligence, present an opportune moment to harness these tools for the benefit of ASD research. The increasing availability of large datasets and computational resources allows for the development of sophisticated models that were not feasible in the past.

6. **Interdisciplinary Collaboration:**
The project is motivated by the belief that interdisciplinary collaboration between clinicians, researchers, and data scientists is essential for addressing the complex challenges associated with ASD. By fostering collaboration, this project aims to bridge the gap between clinical expertise and technological advancements.

7. **Potential to Uncover Biological Mechanisms:**
Machine learning can contribute to our understanding of the underlying biological mechanisms of ASD. By analyzing neuroimaging and genetic data, the project seeks to identify biomarkers and patterns that may offer insights into the neurobiological basis of ASD, opening avenues for targeted research and treatment approaches.

8. **Ethical and Responsible AI Use:**
The motivation includes a commitment to addressing ethical considerations associated with the use of machine learning in healthcare, particularly in sensitive areas like ASD diagnosis. The project aims to explore and implement strategies to ensure responsible and ethical use of machine learning in the context of ASD research.

In essence, the motivation behind this project lies in the potential of machine learning to revolutionize ASD research and diagnosis, contributing to early intervention strategies and personalized support for individuals on the autism spectrum. The goal is to harness technological advancements to make a meaningful impact on the lives of those affected by ASD and their families[5].

1.3 Background

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by a spectrum of challenges in social communication, repetitive behaviors, and restricted interests. It affects individuals across diverse demographic and cultural backgrounds, with varying degrees of severity and symptomatology. The prevalence of ASD has been on the rise globally, emphasizing the need for innovative approaches to understanding, diagnosing, and intervening in this complex disorder.

Traditional methods of ASD diagnosis heavily rely on clinical observations, standardized assessments, and parent/caregiver interviews. While these methods have been invaluable, they often come with limitations, including subjectivity, potential for misdiagnosis, and reliance on observable behaviors. Additionally, the heterogeneity of ASD poses a challenge, as individuals with similar observable symptoms may have distinct underlying characteristics and needs.

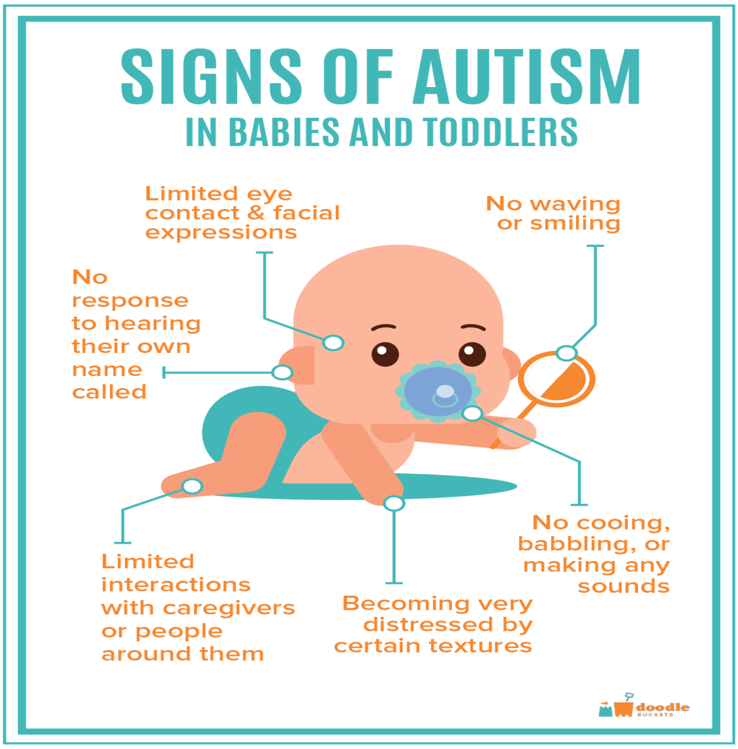
SIGNS OF AUTISM
IN BABIES AND TODDLERS
Limited eye contact & facial expressions
No waving or smiling
No response to hearing their own name called
No cooing, babbling, or making any sounds
Limited interactions with caregivers or people around them
Becoming very distressed by certain textures
doodle BUCKETS

Figure 3: Common symptoms of ASD

The integration of machine learning (ML) into ASD research represents a paradigm shift in the way we approach diagnosis and understanding of the disorder. Machine learning algorithms, fueled by advancements in computational capabilities and data availability, offer the potential to analyze vast amounts of data from diverse sources. This includes behavioral assessments, neuroimaging data, genetic information, and other modalities, allowing for a more comprehensive and objective understanding of ASD.

Machine learning techniques have shown promise in identifying subtle patterns, correlations, and features within datasets that may not be readily apparent through traditional diagnostic methods. By leveraging computational power, ML models aim to improve the accuracy and efficiency of ASD diagnosis, facilitate early detection, and contribute to a deeper understanding of the underlying biological mechanisms associated with ASD.

The background of this project is grounded in the recognition of the limitations of current diagnostic practices, the complexity of ASD, and the potential of machine learning to revolutionize our approach. By exploring the intersection of ASD and machine learning, the project seeks to address these challenges, contribute to the scientific understanding of ASD, and provide practical insights for the development of more effective, personalized interventions for individuals on the autism spectrum.

2. Work Description

This project involves a comprehensive exploration of the application of machine learning (ML) techniques to enhance our understanding of Autism Spectrum Disorder (ASD) and improve diagnostic processes. The work can be broken down into several key components:

1. Literature Review

Conduct an extensive review of existing literature on ASD, machine learning applications in healthcare, and the intersection of these fields. Explore the current state of ASD diagnosis, challenges faced in traditional methods, and the potential contributions of machine learning to address these challenges

2. Data Exploration

The project begins with a thorough exploration of the dataset developed by Dr Fadi Fayez Thabtah [14] to screen autism in toddlers. This involves data cleaning, validation, and preprocessing to ensure the integrity and reliability of the information. 10 behavioral features have been recorded in addition to some individual characteristics that are effective in detecting ASD cases

3. Descriptive Analysis

Initial descriptive analyses provide an overview of the prevalence of autism spectrum disorder in toddlers and the distribution of relevant variables. This stage sets the foundation for subsequent in-depth investigations.

4. Machine Learning Models

To investigate predictive linkages and trends within the dataset, the research makes use of machine learning algorithms such as logistic regression, Support Vector Machines (SVM), and k-Nearest Neighbors (kNN). These models extend the reach of conventional statistical techniques, enabling a more sophisticated comprehension of the interactions between variables.

- **Logistic Regression:** This model is employed to analyze the relationship between the dependent variable (ASD Traits) and independent variables. Logistic regression provides insights into the probability of autism spectrum disorderbased on various factors.
- **Support Vector Machines (SVM):** SVM is utilized for classification tasks, aimingto identify patterns that distinguish between presence and absence of ASD. This model excels in handling non-linear relationships and is well-suited forcomplex datasets.

- **k-Nearest Neighbors (kNN):** For the purposes of classification and clustering, kNN is used. This method assists in identifying patterns and distinctions in the occurrence of autism depending on different variables by taking into account the closest neighbors of data points.
- **Naive Bayes:** Based on the premise of conditional independence of characteristics given the class label, the probabilistic and generative Naive Bayes model functions. This 'naive' assumption frequently works well, especially for text classification and spam filtering applications.
- **Decision Tree:** A decision tree, a type of supervised machine learning model, employs a tree-like structure where each internal node represents a decision based on features, making it an intuitive model for classification and regressiontasks.
- **Random Forest:** Random Forest, an ensemble learning technique based on bagging, enhances predictive accuracy by constructing multiple decision trees. It leverages random subsets of data and features, addressing overfitting concerns and providing valuable feature importance insights.

5. Evaluation and Model Selection:

Define appropriate evaluation metrics to assess the performance of the developed models. Consider metrics such as sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC-ROC) to quantify the models' ability to discriminate between individuals with ASD.

6. Interpretation and Recommendations

Analyze the results obtained from the machine learning models and interpret the findings in the context of ASD diagnosis. Explore the significance of identified features, patterns, and potential biomarkers to contribute to the scientific understanding of ASD.

3. Technical Specification

A strong technical framework is developed in the course of this study to examine the connections between toddler autism spectrum disorder and a number of independent variables. The technical specification calls for the use of machine learning models such as Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and logistic regression. Data cleaning is a preprocessing phase in the overall workflow that guarantees the accuracy and consistency of the data before model analysis.

Data Preprocessing:

Carefully cleansing the data is done before submitting it to machine learning algorithms, this involves-

Handling Missing Data

Identification and treatment of missing values to ensure completeness in the dataset.

Outlier Detection and Treatment

Identifying and dealing with outliers that might distort the analysis's findings.

Ensuring Data Consistency

Verification of data consistency and accuracy to prevent errors in thesubsequent analysis.

Formatting and Standardization

Ensuring uniformity in data format and standardizing variables to facilitate meaningful comparisons.

EDA:

The process of examining and analyzing record sets to recognize patterns, find outliers, and determine the connections between variables is known as exploratory data analysis, or EDA. EDA is typically done as a first step before more formal statistical studies or modeling are done. Exploratory Data Analysis (EDA) is a crucial phase in the data analysis process that involves examining and visualizing data to understand its main characteristics, patterns, and relationships. In the context of Autism Spectrum Disorder (ASD) research with machine learning, EDA helps researchers gain insights into the structure and nature of the data, which is essential for informed feature selection, model development, and interpretation of results.

Machine Learning Models:

Support Vector Machines (SVM)

One of the most widely used supervised learning techniques for both classification and regression issues is support vector machine, or SVM. But it's mostly applied to machine learning classification challenges. In order to make it simple to classify fresh data points in the future, the SVM method seeks to identify the optimal line or decision boundary that may divide n-dimensional space into classes. We refer to this optimal decision boundary as a hyperplane.

SVM selects the extreme vectors and points to aid in the creation of the hyperplane. The technique is referred regarded as a Support Vector Machine since these extreme situations are known as support vectors. A machine learning technique used for classification jobs is SVM. Based on several independent factors, SVM might be utilized in this study's setting to categorize or forecast autism. SVM is well-known for its adaptability to a variety of data sources and is especially useful in addressing non-linear correlations.

k-Nearest Neighbors (kNN)

One of the most basic machine learning algorithms, K-Nearest Neighbor, is based on the supervised learning approach. The K-NN method places the new case in the category most comparable to the existing categories based on its assumption that the new instance and its data are similar to the examples that are already available. The K-NN method classifies a new data point based on similarity after storing all the relevant data. This indicates that the K-NN algorithm can quickly classify newly discovered data into a well-suited category.

Although the K-NN technique is mostly utilized for classification issues, it may also be used for regression. K-NN does not make any assumptions about the underlying data because it is a non-parametric method.

A machine learning approach called kNN is employed for both regression and classification problems. Based on the closeness of data points in the feature space, kNN might be used in this study's setting to find trends or similarities in autism spectrum disorder. This technique can be useful for deciphering the data's structure since it is sensitive to local patterns.

Logistic Regression

One of the most widely used machine learning algorithms, within the category of supervised learning, is logistic regression. With a given collection of independent factors, it is used to predict the categorical dependent variable. With logistic regression, the result of a categorical dependent variable is predicted. As a result, a discrete or category value must be the result. Instead of providing the precise values, which are 0 and 1, it provides the probabilistic values, which fall between 0 and 1. It can be either Yes or No, 0 or 1, true or False, etc. With the exception of how they are applied, logistic regression and linear regression are very similar. While logistic regression is used to solve classification difficulties, linear regression is used to solve regression problems. The logistic function curve shows the probability of several things, such whether the cells are malignant or not, if a mouse is fat depending on its weight, etc. Using both continuous and discrete datasets, logistic regression is a powerful machine learning technique that can generate probabilities and categorize new data. Different forms of data may be utilized to categorize observations using logistic regression, which also makes it simple to identify the most useful factors for the classification.

A popular statistical technique for binary classification problems is logistic regression. It simulates the likelihood that an event will occur. Logistic regression may be utilized in this study to forecast a child's probability of developing autism depending on certain independent factors.

Naive Bayes

The Naïve Bayes algorithm is a supervised learning technique that solves classification issues. It is based on the Bayes theorem. Its primary use is in text categorization, where a high-dimensional training dataset is employed. One of the most straightforward and efficient classification algorithms, the Naïve Bayes classifier aids in the rapid development of machine learning models with rapid prediction capabilities. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur. Among the most well-known applications of the Naïve Bayes algorithm include article classification, sentiment analysis, and spam filtering. Based on the premise of conditional independence of characteristics given the class label, the probabilistic and generative Naive Bayes model functions. This 'naive' assumption frequently works well, especially for text classification and spam filtering applications.

Decision Tree

Although decision trees are a supervised learning approach, they are mostly employed to solve classification issues. However, they may also be used to solve regression problems. This classifier is tree-structured, with internal nodes standing in for dataset attributes, branches for decision rules, and leaf nodes for each outcome. The Decision Node and the Leaf Node are the two nodes that make up a decision tree. While leaf nodes represent the result of decisions and do not have any more branches, decision nodes are used to make any kind of decision and have numerous branches. The characteristics of the provided dataset are used to inform the decisions or the test. A decision tree, a type of supervised machine learning model, employs a tree- like structure where each internal node represents a decision based on features, making it an intuitive model for classification and regression tasks.

Random Forest

Among the supervised learning methods is the well-known machine learning algorithm Random Forest. It may be applied to ML issues involving both classification and regression. Its foundation is the idea of ensemble learning, which is the act of merging several classifiers to solve a challenging issue and enhance the model's functionality. According to its name, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Rather of depending on a single decision tree, the random forest forecasts the outcome based on the majority vote of projections from each tree. Random Forest, an ensemble learning technique based on bagging, enhances predictive accuracy by constructing multiple decision trees. It leverages random subsets of data and features, addressing overfitting concerns and providing valuable feature importance insights

Evaluation metrics:

An evaluation matrix is a structured table or framework used to systematically assess and compare different aspects of a system, process, or model. It is designed to help evaluate the performance or effectiveness of various components or alternatives based on specific criteria. Accuracy, precision, recall, F1 score, and AUC-ROC (Area Under the Receiver Operating Characteristic Curve) are commonly used metrics in machine learning for evaluating the performance of classification models. Here's a brief explanation of each:

Accuracy:

Accuracy is the most straightforward metric and represents the ratio of correctly predicted instances to the total instances in the dataset.

Precision:

Precision is the ratio of correctly predicted positive observations to the total predicted positives. It focuses on the accuracy of positive predictions.

Recall (Sensitivity):

Recall, also known as sensitivity or true positive rate, measures the ability of the model to capture all the positive instances.

F1 Score:

F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially when there is an uneven class distribution.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve):

AUC-ROC is a metric used to assess the performance of a classification model at various thresholds. The ROC curve is a graphical representation of the true positive rate against the false positive rate. AUC-ROC measures the area under this curve, where a higher AUC-ROC indicates better model performance. It ranges from 0 to 1, with 0.5 representing a model with no discriminatory power and 1 representing a perfect model.

Confusion Matrix

A confusion matrix is a table that is often used to evaluate the performance of a classification algorithm. It provides a clear representation of the model's predictions by comparing them to the actual class labels. The confusion matrix is particularly useful when dealing with binary or multiclass classification problems.

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 4: Evaluation metrics formulas

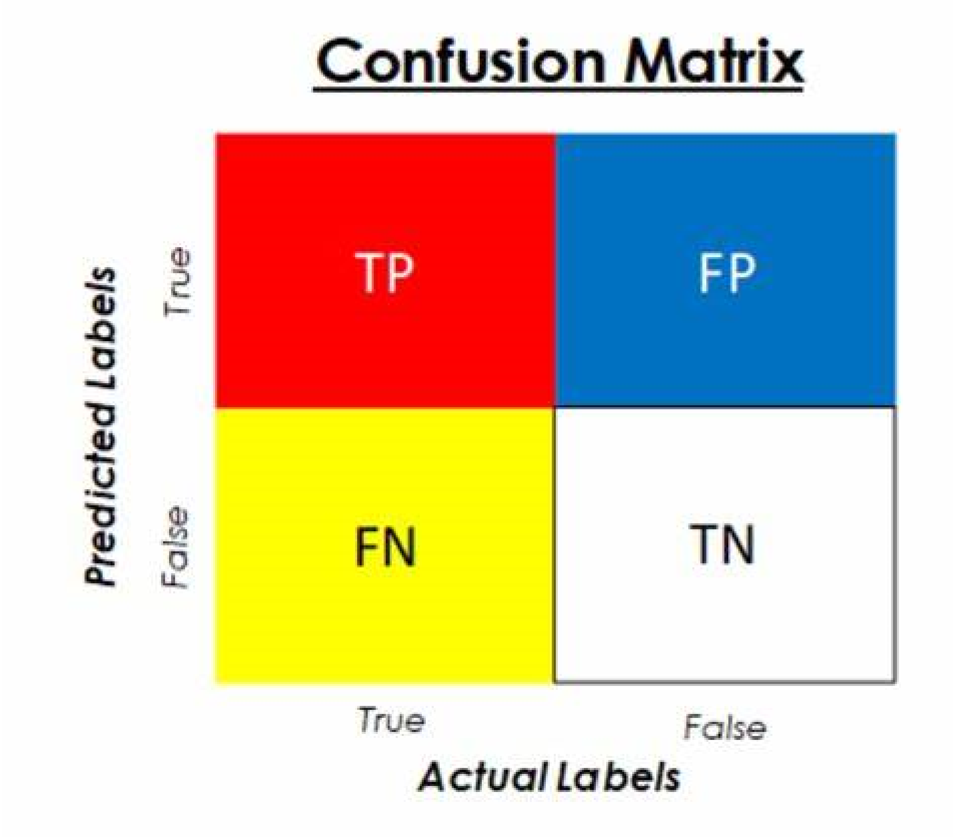

Figure 5: Confusion Matrix

4. Literature Survey

Although concerns about children development milestones are mostly reported by parents of children with ASD within their first year, these children are rarely diagnosed before the age of 4 years [3, 6]. Early identification is key as, if followed by intensive early intervention, will enable these kids to acquire the necessary skills, improve the core behavioral symptoms, and there is a chance that they will grow out of the diagnosis or at least loose many of the autistic traits [7]. A significant number of studies investigated the potential for early intervention in helping kids with this neurodevelopmental condition. In this context, [8] conducted a systematic review that highlighted the large number of studies, which are over 83% of the published literature since 2010, reflecting the increased interest among researchers in this area.

Autism cannot be identified using conventional clinical methods such as blood tests. Instead, ASD screening is a crucial phase, and it is the process of determining the autistic symptoms of an individual. Many screening tools have emerged overtime, they include direct observations, questionnaires and interviews and they should be performed by specialists. However, these tools are lengthy, costly, and low-and-middle-income countries have shortage of mental health clinicians. This constitutes a major barrier for obtaining an early diagnosis and accessing intervention services [9]. Therefore, a viable screening instrument that is cost effective, available and less time consuming to identify the risk of ASD at a preliminary stage is highly needed.

There is a growing interest among researchers in the early screening and intervention for young children at risk of ASD [8] and in the use of machine learning and intelligent methods for autism screening and detection [9, 10]. In this context, [11] investigated fuzzy data mining models to detect autism features for both cases and control groups between 4 and 11 years. [5] developed an effective prediction model by merging Random Forest-CART and Random Forest-Id3 and complemented their work by developing a mobile application based on the proposed model. [12] developed and tested an Artificial Neural Network (ANN) for diagnosing ASD, and the test data evaluation showed that ANN model was able to diagnose ASD with 100% accuracy. [13] used machine learning to investigate the accuracy and reliability of the Q-CHAT method in classifying autistic kids. The authors used three different ML algorithms and found that the Support Vector Machine was the most effective. Similarly, [10] proposed a new machine learning method which is the “Rules-Machine Learning”. This method detects the ASD traits in cases and offers rules that can be used by domain experts to understand the reasons behind the classification. The authors compared this method with other classifiers and found that it leads to higher predictive accuracy, sensitivity, and specificity than those of other models such as bagging, decision trees and rule induction. In line with these studies, this paper contributes to this area by proposing a toddler screening prediction model for ASD traits.

5. Methodology

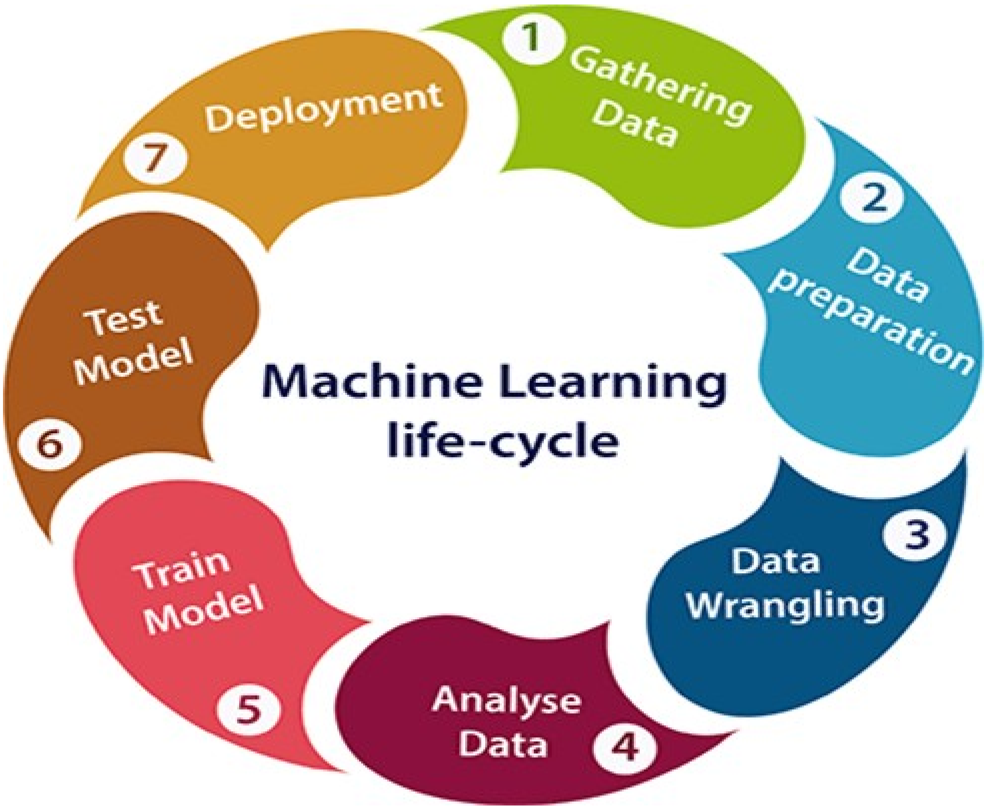

Figure 6: Machine learning life cycle model

1. Gathering Data

The initial phase of the machine learning life cycle is data collection. This step's objective is to locate and acquire every issue pertaining to data. We must identify the numerous data sources in this stage since information may be gathered from a variety of places, including files, databases, the internet, and mobile devices. It is among the life cycle's most crucial stages. The output's efficiency will depend on the volume and caliber of data gathered. The forecast will be more accurate the more data there is [16].

The following tasks are part of this step:

Determine many sources of data. Gather information

Combine the information gathered from various sources.

2. Data preparation

After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepareit to use in our machine learning training. In this step, first, we put all data together, and then randomize the ordering ofdata [16].

This step can be further divided into two processes:

Data exploration:

It is used to understand the nature of data that we have to work with. Weneed to understand the characteristics, format, and quality of data. A better understanding of data leads to an effective outcome. In this, we findCorrelations, general trends, and outliers.

Data pre-processing:

Now the next step is preprocessing of data for its analysis [16].

3. Data Wrangling

Cleaning and transforming unprocessed data into a format that can be used is known as data wrangling. In order to prepare the data for analysis in the following phase, it is necessary to clean it, choose the variable to utilize, and format it correctly. It is among the most crucial actions in the entire procedure. In order to solve the quality concerns, data cleaning is necessary. It is not necessary that data we have collected is always of our use as some of the data may not be useful. In real-world applications, collected data may have various issues, including:

Missing Values,Duplicate data , Invalid data, Noise

So, we use various filtering techniques to clean the data.

It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome [16].

4. Data Analysis

Now the cleaned and prepared data is passed on to the analysis step. This step involves:

Selection of analytical techniquesBuilding models

Review the result

This step's objective is to create a machine learning model that will study the data using a variety of analytical methods and evaluate the results. First, the nature of the issue must be determined. Next, several machine learning approaches, including classification, regression, cluster analysis, association, and others, must be chosen. The model must then be constructed using provided data and evaluated. Hence, in this step, we take the data and use machine learning algorithms tobuild the model [16].

5. Train Model

Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem. We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features [16].

6. Test Model

Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it. Testing the model determines the percentage accuracy of the model as per therequirement of project or problem [16].

7. Deployment

The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system. If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system. But before deploying the project, we will check whether it is improving its performance using available data or not. The deployment phase is like making the final report for a project [16].

Based on above-mentioned process, we choose to proceed with Logistic Regression for model building process:

A logistic function, sometimes called a sigmoid function, is used in logistic regression, a supervised machine learning technique, to provide a probability value between 0 and 1 based on inputs that are independent variables. As an illustration, there are two classes: Class 0 and Class 1. An input is classified as Class 1 or Class 0 if the logistic function value for it is larger than the threshold value of 0.5. Since it is a continuation of linear regression and is mostly applied to classification issues, it is known as regression. The output of linear regression is a continuous number that can be anything, but the prediction of probability is the difference between logistic regression and linear regression[17].

With a given collection of independent factors, it is used to predict the categorical dependent variable.

With logistic regression, the result of a categorical dependent variable is predicted. As a result, a discrete or category value must be the result.
Instead of providing the precise values, which are 0 and 1, it provides the probabilistic values, which fall between 0 and 1. It can be either Yes or No, 0 or 1, true or False, etc. Except for how they are applied, logistic regression and linear regression are very similar. While logistic regression is used to solve classification difficulties, linear regression is used to solve regression problems.
In logistic regression, we fit a "S" shaped logistic function, which predicts two maximum values, rather than a regression line[17].

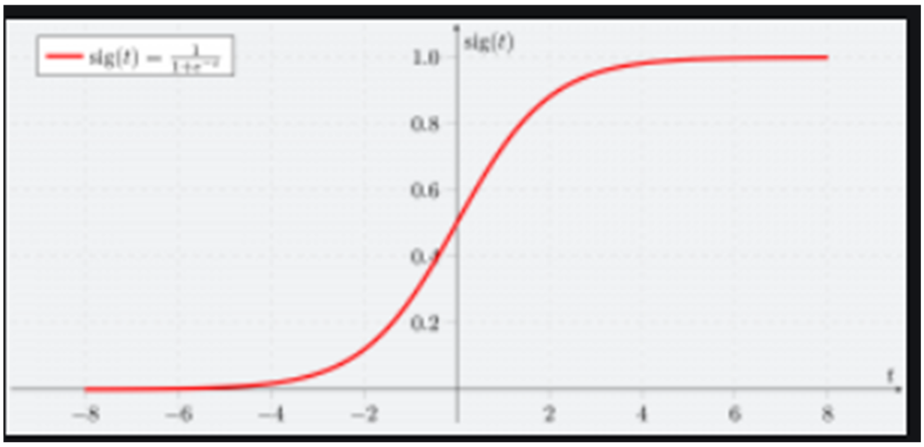

Figure 7:Sigmoid function

Logistic Function (Sigmoid Function): A mathematical function that converts expected values into probabilities is the sigmoid function. It converts any real number between 0 and 1 into another value. The logistic regression's result must be between 0 and 1, and as it cannot be greater than this, it takes the shape of a "S" curve.
The logistic or sigmoid function is another name for the S-form curve.
The idea of the threshold value, which indicates the likelihood of either 0 or 1, is used in logistic regression. For example, numbers above the threshold tend toward one, whereas those below the threshold tend toward zero[17].

Types of Logistic Regression
Based on the categories, Logistic Regression can be classified into three types:

Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”
Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”[17].

Applying steps in logistic regression modeling
The following are the steps involved in logistic regression modeling:

Define the problem: Identify the dependent variable and independent variables and determine if the problem is a binary classification problem.
Data preparation: Clean and preprocess the data, and make sure the data is suitable for logistic regression modeling.
Exploratory Data Analysis (EDA): Visualize the relationships between the dependent and independent variables, and identify any outliers or anomalies in the data.
Feature Selection: Choose the independent variables that have a significant relationship with the dependent variable, and remove any redundant or

irrelevant features.

Model Building: Train the logistic regression model on the selected independent variables and estimate the coefficients of the model.

Model Evaluation: Evaluate the performance of the logistic regression model using appropriate metrics such as accuracy, precision, recall, F1-score, or AUC-ROC.

Model improvement: Based on the results of the evaluation, fine-tune the model by adjusting the independent variables, adding new features, or using regularization techniques to reduce overfitting[17].

6. Output

About the dataset:

The dataset included in this work was created by Dr. Fadi Fayez Thabtah [14] to screen for autism in children, and it was retrieved from the Kaggle website. In addition to a few individual traits that are useful in identifying ASD patients, this dataset contains 10 behavioral elements.

The dataset's numerous properties are compiled in Table 1.

Variable	Type	Description
A1:Response_to_name	Binary(0,1)	Does your child look at you when you call his/her name?
A2:Eye_contact	Binary(0,1)	How easy is it for you to get eye contact with your child?
A3:Point_to_objects	Binary(0,1)	Does your child point to indicate that s/he wants something?
A4: Sharing_interest	Binary(0,1)	Does your child point to share interest with you?
A5: Pretend_play	Binary(0,1)	Does your child pretend?
A6: Follow_looking	Binary(0,1)	Does your child follow where you are looking?
A7:Confort_someone	Binary(0,1)	does your child show signs of wanting to comfort someone upset?
A8: First_words	Binary(0,1)	Description of child first words

A9: Simple_gesture	Binary(0,1)	Does your child use simple gestures?
A10:Stare_at_nothing	Binary(0,1)	Does your child stare at nothing with no apparent purpose?
Age	Number	Toddlers (months)
Score by Q-chat-10	Number	1-10 (Less than or equal 3 no ASD traits; > 3 ASD traits
Sex	Character	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean(Y/N)	Whether the case was born with jaundice
Family member with ASD history	Boolean(Y/N)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician, etc.
Class variable (ASD)	String	ASD traits or No ASD traits (Yes / No)

Table 1: Dataset Description (source: [14])

The available responses to the questions in items A1 through A10 are Always, Usually, Occasionally, Rarely, and Never. In relation to questions A1 through A9, a score of 1 is allocated when the response is Sometimes, Rarely, or Never, while a score of 0 is awarded when the response is Always or Usually. On the other hand, for question 10, 1 is given if the response was Always, Usually, or Sometimes; otherwise, 0 is given. The age of toddlers is between 1 to 3 years.

Types of Graph plotted: Bar graph, Pie plot, Strip plot, Pair plot, Violin plot, Heatmap, Histogram, Count plot, KDE plot, Line plot, Scatter plot.

Results

1. Accuracies of different models

Model	Accuracy
Logistic Regression	95
SVM	94
Naive Bayes	91
k-NN	93
Decision Tree	90
Random Forest	95

Table 2: Algorithm and Accuracies

2. Logistic Regression

Parameters	Logistic	GridSearchCV	RandomSearchCV
Accuracy	95	95	95
Precision	94	94	94
Recall	95	95	95
F1 score	95	95	95

Table 3:Logistic Regression Result

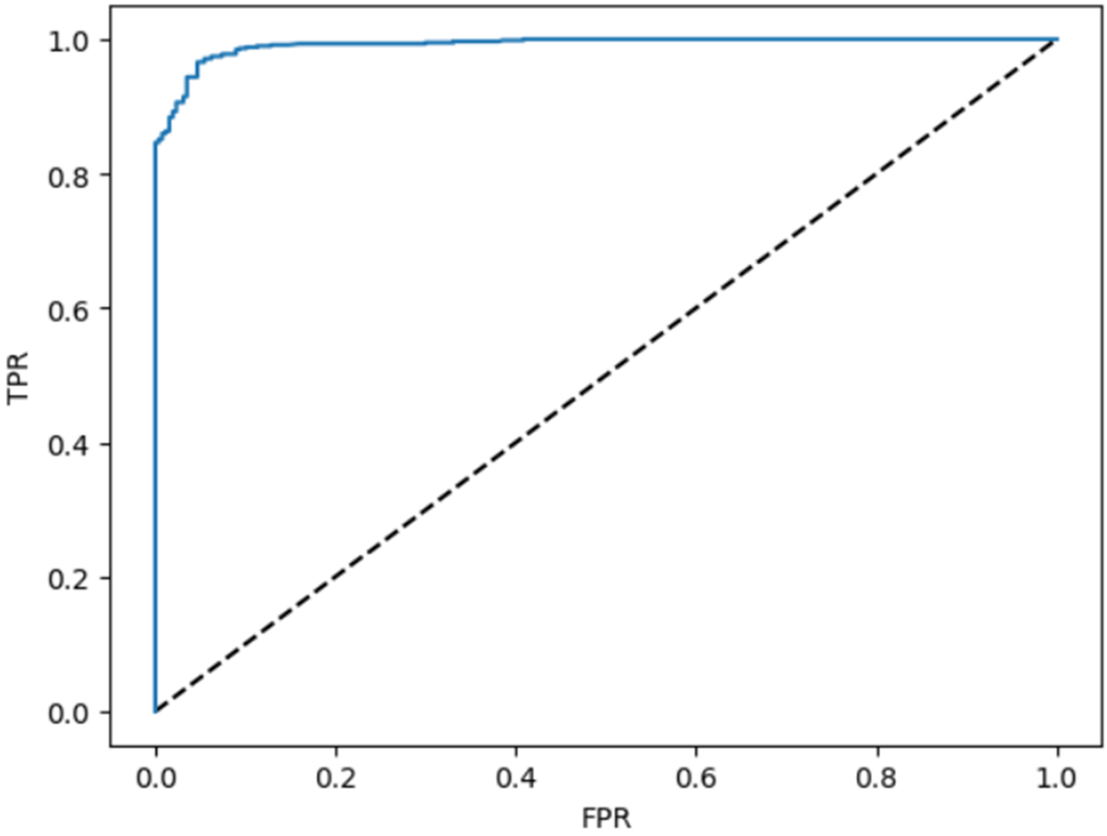

Figure 8: AUC-ROC curve

7. Conclusion

There is an immediate need to consider creating and implementing a quick, easy-to-use, and affordable autism screening tool because the number of cases receiving a diagnosis of autism is rising, and there are numerous obstacles standing in the way of receiving a timely screening and a diagnosis that would enable the children to receive early intervention services. This was the goal of the study that demonstrated the machine learning algorithms' ability to predict autistic features. The dataset used to train the model included information on gender, age, ethnicity, some aspects of family history, and the relationship to the person who completed the screening, in addition to ten behavioral features that reflect developmental milestones and are related to social and communicative behaviors. After training multiple classifiers, one model was shortlisted as candidates for providing exceptionally good accuracy, sensitivity, and specificity. Although the three models' performances are excellent, that is Logistic Regression, and therefore we propose to implement it. This study showed promising results for ASD screening, and the proposed model can be complemented by developing a mobile application that will add to the convenience and accessibility of the screening method.

8. Reference

1. CDC. Reports, (2020) "AUTISM AND DEVELOPMENTAL DISABILITIES MONITORING (ADDM) NETWORK," ed.
2. V. Jagan and A. Sathiyaseelan, (2016) "Early intervention and diagnosis of autism," Indian Journal of Health and Wellbeing, vol. 7, no. 12, pp. 1144-1148.
3. L. E. K. Achenie, A. Scarpa, R. S. Factor, T. Wang, D. L. Robins, and D. S. McCrickard, (2019) "A Machine Learning Strategy for Autism Screening in Toddlers," Journal of Developmental and Behavioral Pediatrics, vol. 40, no. 5, p. 369, doi: 10.1097/DBP.0000000000000668.
4. P. S. Carbone et al., (2020) "Primary Care Autism Screening and Later Autism Diagnosis," Pediatrics, vol. 146, no. 2, doi: 10.1542/peds.2019-2314.
5. K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi and M. N. Islam, (2019) "A Machine Learning Approach to Predict Autism Spectrum Disorder," in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE): IEEE, pp. 1-6.
6. S. Broder Fingert et al., (2019) "Implementing systems-based innovations to improve access to early screening, diagnosis, and treatment services for children with autism spectrum disorder: An Autism Spectrum Disorder Pediatric, Early Detection, Engagement, and Services network study," Autism : the international journal of research and practice, vol. 23, no. 3, pp. 653-664, doi: 10.1177/1362361318766238.
7. C. Kamuk, C. Cantio, and N. Bilenberg, (2017) "Early screening for autism spectrum disorder," European Psychiatry, vol. 41, no. Supplement, pp. S131-S132, doi: 10.1016/j.eurpsy.2017.01.1948. Computer Science & Information Technology (CS & IT).
8. L. French and E. M. M. Kennedy, (2018) "Annual Research Review: Early intervention for infants and young children with, or at-risk of, autism spectrum disorder: a systematic review," Journal of Child Psychology and Psychiatry, vol. 59, no. 4, pp. 444-456, doi: 10.1111/jcpp.12828.
9. B. Wingfield et al., (2020) "A predictive model for paediatric autism screening," Health informatics journal, p. 1460458219887823, doi: 10.1177/1460458219887823.
10. F. Thabtah and D. Peebles, (2020) "A new machine learning model based on induction of rules for autism detection," Health informatics journal, vol. 26, no. 1, pp. 264-286, doi: 10.1177/1460458218824711.
11. M. Al-diabat, (2018) "Fuzzy data mining for autism classification of children," International Journal of Advanced Computer Science and

Applications, vol. 9, no. 7, pp. 11-17, doi: 10.14569/IJACSA.2018.090702.

12. I. M. Nasser, M. O. Al-Shawwa, and S. S. Abu-Naser, (2019) "Artificial Neural Network for Diagnose Autism Spectrum Disorder " International Journal of Academic Information Systems Research (IJAISR), vol. 3, no. 2, pp. 27-32.
13. T. Gennaro, C. Giovanni, D. P. Davide, and A. Stefania, (2020) "Use of Machine Learning to Investigate The Quantitative Checklist For Autism in Toddlers (QCHAT) Towards Early Autism Screening," ed.
14. F. Thabtah, (2020) "Autism screening data for toddlers." https://www.kaggle.com/fabdelja/autism screening-for-toddlers (accessed.
15. R. Loomes, L. Hull, and W. Polmear, (2017) "What is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis " Journal of the American Academy of Child & Adolescent Psychiatry.
16. https://www.javatpoint.com/machine-learning-life-cycle
17. https://www.geeksforgeeks.org/understanding-logistic-regression/