A Mini Project report

On

AUTISM SPECTRUM DISORDER CLASSIFICATION AND PREDICTION

A dissertation Submitted in partial fulfillment of the

requirement for the

Award of the degree of

BACHELOR OF TECHNOLOGY

In

ELECTRONICS AND COMMUNICATION ENGINEERING

By

M.GANESH	**21N31A04D8**
P. GEORGE	**21N31A04G0**
R. RAJASHEKAR	**21N31A04J7**

Under the esteemed guidance of

Dr. R CHINNA RAO

Associate Professor

DEPARTMENT OF ELECTRONICS & COMMUICATION ENGINEERING

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

Autonomous Institution – UGC, Govt. of India

(Affiliated to JNTU, Hyderabad, Approved by AICTE - Accredited by NBA & NAAC – 'A' Grade)

Maisammaguda, Dhulapally (Post Via Hakimpet), Secunderabad – 500100

2021-2025

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

2021-2025

CERTIFICATE

This is to certify that the **MINI PROJECT** work entitled "**AUTISM SPECTRUM DISORDER CLASSIFICATION AND PREDICTION**" is carried out by **M.GANESH (21N31A04D8),P.GEORGE (21N31A04G0), R.RAJASHEKAR (21N31A04J7)**,in partial fulfillment for the award ofdegree of **BACHELOR OF TECHNOLOGY** in Electronics and communication Engineering, Jawaharlal Nehru Technological University ,Hyderabad during the academic year 2024-2025.

INTERNAL GUIDE	**HEAD OF THE DEPARTMENT**
Dr R CHINNA RAO	**Dr K MALLIKARJUNA LINGAM**
ASSOCIATE PROFESSOR	**PROFESSOR & HOD**

EXTERNAL EXAMINER

DECLARARTION

We hereby declare that the minor project entitled "**Autism Spectrum Disorder Classification and Prediction"** submitted in partial fulfillment of the requirements for the award of degree of **Bachelor of Technology (B. Tech)** in **Electronics and Communication Engineering (ECE).** This dissertation is our original work and the project has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles and no part of it has been published or sent for the publication at the time of submission.

BY

M. GANESH 21N31A04D8

P. GEORGE 21N31A04G0

R. RAJASHEKAR 21N31A04J7

ACKNOWLEDGMENTS

First and foremost, we would like to thank God for giving strength to complete this work. We feel ourselves honoured and privileged to place our warm salutation to our college **Malla Reddy College of Engineering and Technology (UGC-Autonomous)** and our director **Dr. VSK Reddy,** principal **Dr. S. Srinivasa Rao** who gave us the opportunity to have experience in engineering and profound technical knowledge. We express our heartiest thanks to **Prof P. Sanjeeva Reddy**, Director of the Department of Electronics and Communication Engineering and **Dr. K. Mallikarjuna Lingam**, Head of the Department of Electronics and communication engineering for encouraging us in every aspect of our project and helping us realize our full potential. We would like to thank our Project Coordinator **Ms D. Asha**, Assistant Professor, Department of Electronics and Communication Engineering for their valuable guidance and encouragement during my project work. We would like to thank our internal guide **Dr. R. Chinna Rao**, Associate Professor, Department of Electronics and Communication Engineering for his regular guidance and constant encouragement. We are extremely grateful to his valuable suggestions and unflinching cooperation throughout project work. We would like to thank **Dr. T. Venugopal**, Dean of Academics, who in spite of being busy with his duties took time to guide and keep us on the correct path. We would also like to thank all the supporting staff of the Department of Electronics and Communication Engineering and all other departments who have been helpful directly or indirectly in making our project a success. We are extremely grateful to our parents for their blessings and prayers for the completion of our project that gave us strength to do our project.

BY

M.GANESH	**21N31A04D8**
P. GEORGE	**21N31A04G0**
R. RAJASHEKAR	**21N31A04J7**

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Symbol	Symbol Name	Meaning / definition
+	Addition	Indicates addition or presence of a feature.
-	Subtraction	Indicates subtraction or absence of a feature.
%	Percentage	Represents performance metrics or prevalence.
=	Equals	Represents equality, used in equations or assignments.
$\geq$	Greater than or equal to	Indicates a value is greater than or equal to a threshold.
$\leq$	Less than or equal to	Indicates a value is less than or equal to a threshold.
/	Division or ratio	Represents division or a ratio between two values.
[ ]	Range brackets	Denotes a range of values, such as for normalization.
$\sum$	Summation	Represents summation over a set of values.
μ	Mean	Represents the average value in statistical calculations.
σ	Standard deviation	Represents the standard deviation, used in standardization.
n	Sample size	Represents the sample size or number of iterations.

LIST OF ACRONYMS

1. **ASD –** Autism Spectrum Disorder

2. **CDC –** Centers for Disease Control and Prevention

3. **DSM-5** – Diagnostic and Statistical Manual of Mental Disorders, 5th Edition

4. **ICD-10** – International Classification of Diseases, 10th Revision

5. **ABA –** Applied Behavior Analysis

6. **ADOS** – Autism Diagnostic Observation Schedule

7. **ADI-R** – Autism Diagnostic Interview-Revised

8. **PCA –** Principal Component Analysis

9. **ROC –** Receiver Operating Characteristic

10. **AUC** – Area Under the Curve

11. **SVM** – Support Vector Machines

12. **ML –** Machine Learning

13. **UAT** – User Acceptance Testing

14. **AWS** – Amazon Web Services

15. **API –** Application Programming Interface

16. **EMF** – Enhanced Meta Format

ABSTRACT

This project aims to develop a machine learning model to classify and predict the likelihood of Autism Spectrum Disorder (ASD) in individuals based on various behavioural, developmental, and demographic features. ASD is a neurodevelopmental condition characterized by challenges in social interaction, communication, and repetitive behaviours. Early detection is critical in providing effective interventions that can improve the quality of life for those affected.

The dataset used includes screening questions that assess social interaction, communication patterns, repetitive behaviours, sensory sensitivities, developmental milestones, and responses to treatments. Additional features such as age, gender are incorporated to enhance the predictive capability of the model.

A Random Forest classifier has been chosen as the core algorithm due to its robustness and ability to handle complex, non-linear relationships within the data. The model is trained on pre-processed and standardized data to ensure accurate predictions. The system allows users to input responses to key questions about ASD-related behaviours, and based on these inputs, it calculates the probability of an ASD diagnosis.

The proposed solution provides a reliable and user-friendly tool for early ASD screening, offering a decision support system for healthcare professionals and caregivers. The model's performance is evaluated using accuracy metrics, and the tool is designed to be scalable for further enhancements, such as integrating additional features or adjusting for population-specific variations. This approach has the potential to aid in early detection, making timely interventions possible, which can significantly improve outcomes for individuals on the autism spectrum.

CHAPTER 1

INTRODUCTION

1.1 EARLY DETECTION AND DIAGNOSIS

1.1.1 Impact on Intervention

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects communication, social interaction, and behaviour. Early detection is critical because it allows for timely interventions, which can significantly improve the developmental outcomes of children with ASD. Machine learning models can help automate this detection process by analysing patterns in behavioural, sensory, and demographic data, enabling earlier diagnoses and, consequently, earlier intervention.

1.1.2 Reducing Diagnostic Delays

One of the challenges in ASD diagnosis is the time it takes for caregivers and healthcare professionals to identify symptoms and confirm a diagnosis. By leveraging machine learning, this project aims to reduce the time needed for diagnosis, making the process more efficient and accessible. Automated systems can quickly assess behavioural and developmental indicators, identifying patterns that may not be immediately apparent through manual observation.

1.2 IMPROVING ACCURACY AND CONSISTENCY

1.2.1 Overcoming Human Limitations

Traditional ASD diagnostic tools rely on the subjective assessments of medical professionals, which can vary depending on the experience and expertise of the practitioner. Machine learning algorithms, such as Random Forest classifiers, offer more consistent and objective decision-making. They are not subject to human bias and can analyse large datasets in a way that is difficult for humans.

1.2.2 Handling Complex Datasets

ASD symptoms vary widely between individuals, making diagnosis a challenge. This project applies machine learning algorithms that are capable of handling complex, multi-dimensional data and identifying hidden relationships between features such as social interaction patterns, communication difficulties, and repetitive behaviours. By training on a large dataset, the model can learn from diverse cases, improving its ability to generalize across different populations.

1.3 ACCESSIBILITY AND SCALABILITY

1.3.1 Global Applicability

Machine learning-based ASD diagnostic tools can be easily scaled and implemented across various healthcare settings, including remote or under-resourced areas where access to trained specialists is limited. This project could provide a framework for developing applications that can be used by caregivers and professionals around the world to assess the likelihood of ASD based on standardized questionnaires and inputs.

1.3.2 User-Friendly Tool

The project aims to create a user-friendly tool where users can input their responses to screening questions related to ASD. This increases accessibility for both caregivers and healthcare providers, making it easier to integrate into routine developmental checkups. The use of a simple questionnaire interface also allows non-experts to use the tool to screen for potential signs of ASD.

1.4 CUSTOMIZATION FOR SPECIFIC POPULATIONS

1.4.1 Cultural and Demographic Variations

ASD manifests differently in individuals based on cultural, social, and demographic contexts. By including features such as age, gender, ethnicity, and family history, the project provides a more tailored diagnostic tool that can adjust its predictions based on these important factors. This adaptability ensures that the system remains relevant across different populations, improving its accuracy and fairness.

1.4.2 Customizable for Different Age Groups

The project also allows customization for different age groups, recognizing that ASD symptoms in children, adolescents, and adults may present differently. This flexibility enhances the system's accuracy across a wider range of users, making it a versatile tool for screening ASD in both pediatric and adult populations.

1.5 ADVANCEMENT OF RESEARCH AND HEALTHCARE

1.5.1 Contributing to Research

This project not only aims to improve ASD diagnosis but also contributes to the growing body of research that applies machine learning to healthcare. The insights gained from analysing patterns in the data can help researchers understand how ASD manifests in different populations and identify the most important predictors of the disorder.

1.5.2 Support for Healthcare Providers

The predictive model developed in this project can serve as a decision-support tool for healthcare providers, augmenting their expertise with data-driven insights. The model’s ability to provide probability scores for each diagnosis allows healthcare providers to make more informed decisions and focus their attention on high-risk individuals.

1.6 ECONOMIC AND SOCIAL BENEFITS

1.6.1 Cost-Effectiveness

By automating the screening process, this project reduces the costs associated with traditional, in-person diagnostic assessments. This makes ASD screening more affordable and accessible, particularly in regions where healthcare resources are limited. Early detection also reduces long-term healthcare costs by enabling earlier intervention, which can prevent more severe developmental delays.

1.6.2 Improving Quality of Life

Early and accurate diagnosis can significantly enhance the quality of life for individuals with ASD. By providing a tool that facilitates early intervention, this project helps mitigate the challenges associated with late diagnoses, such as difficulties in socialization, education, and independent living. The potential for improving long-term outcomes makes this project highly significant for individuals, families, and communities.

1.7 AUTISM SPECTRUM DISORDER (ASD)

Autism Spectrum Disorder (ASD) is a complex and lifelong neurodevelopmental condition that affects an individual's social interactions, communication skills, and behavior. It is termed a "spectrum" disorder because the symptoms and severity can vary significantly from person to person. The main characteristics of ASD include difficulties with social communication and interaction, as well as restricted, repetitive behaviours, interests, or activities. These challenges can interfere with daily functioning and may affect education, employment, and relationships.

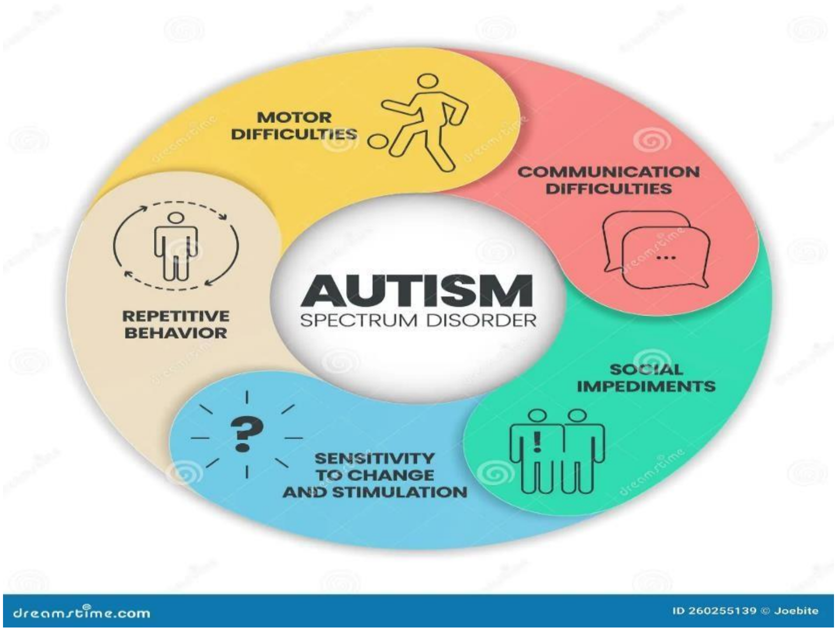


Figure 1.1 : Autism Spectrum Disorder

1.7.1 PREVALENCE AND IMPACT

ASD affects individuals globally, and its prevalence has increased in recent decades due to better awareness, improved diagnostic criteria, and broader screening programs. According to the Centres for Disease Control and Prevention (CDC), approximately 1 in 54 children in the

United States has been identified with ASD. This prevalence underscores the importance of early detection and intervention. Early diagnosis allows children to access therapies and support that can help them develop crucial social, communication, and behavioural skills.

1.7.2 SYMPTOMS AND DIAGNOSIS

ASD typically presents symptoms early in life, often before the age of three, though it can sometimes be diagnosed later in life. Common symptoms include:

- Social Communication Deficits: Individuals with ASD may struggle to maintain eye contact, engage in typical back-and-forth conversations, or understand social cues like facial expressions or body language.
- Repetitive Behaviours and Restricted Interests: Many individuals with ASD engage in repetitive movements, routines, or behaviours (e.g., hand-flapping, insistence on sameness) and may have intense focus on specific topics or interests.
- Sensory Sensitivities: Some individuals may be hypersensitive to sensory inputs such as lights, sounds, textures, or smells.

Diagnosis of ASD is based on the observation of behavior and developmental history. It typically involves multiple evaluations, including input from parents, caregivers, educators, and healthcare professionals. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5), used by clinicians to diagnose ASD, outlines specific criteria focusing on social interaction, communication challenges, and patterns of behavior.

1.7.3 CAUSES AND RISK FACTORS

The exact causes of ASD are not fully understood, but research points to a combination of genetic and environmental factors. Genetic mutations, familial history of autism, and certain environmental influences during prenatal development are believed to increase the risk of developing ASD. There is no known single cause, and research continues to explore how these factors interact.

1.7.4 MANAGEMENT AND INTERVENTION

While there is no cure for ASD, early intervention can make a significant difference. Treatment and support strategies include:

- Behavioral Therapy: Approaches like Applied Behavior Analysis (ABA) focus on improving specific behaviors, social skills, and learning abilities.

- Speech and Occupational Therapy: These interventions help individuals with communication and daily living skills.
- Educational Support: Tailored educational programs and learning accommodations can greatly enhance the academic progress of children with ASD.
- Medications: While no medications treat the core symptoms of ASD, some are used to manage co-occurring issues like anxiety, hyperactivity, or irritability.

1.7.5 LIFELONG PERSPECTIVE

Although ASD is typically diagnosed in early childhood, its effects are lifelong. Many adults with ASD live independently, pursue education and employment, and maintain relationships, though some may need ongoing support. The availability of community resources, social services, and continued research in ASD helps promote a better quality of life for individuals on the spectrum.

In conclusion, ASD is a multifaceted condition that requires a personalized approach to diagnosis and intervention. Advances in research and awareness continue to improve outcomes for individuals with ASD, emphasizing the importance of early detection and comprehensive support throughout life.

1.8 CHALLENGES IN DIAGNOSING AUTISM SPECTRUM DISORDER (ASD)

1.8.1 VARIABILITY OF SYMPTOMS

- **Spectrum Nature:** ASD encompasses a wide range of symptoms and severity levels, leading to variability in presentations.
- **Individual Differences:** Each person with ASD may exhibit different symptoms, making it challenging to establish a clear diagnostic criterion.
- **Comorbid Conditions:** Many individuals with ASD have comorbid disorders (e.g., anxiety, ADHD, intellectual disabilities), complicating the diagnosis and requiring comprehensive assessments.

Table 1.1: Variability in Symptoms

VARIABILITY IN SYMPTOMS	DESCRIPTION
Social interaction	Individuals with ASD may exhibit varying levels of social engagement, from complete withdrawal to active participation, impacting diagnosis.
Communication Skills	Communication abilities can range from non-verbal to highly verbal, with some individuals using alternative methods (e.g., sign language, assistive technology).
Repetitive Behaviors	The type and intensity of repetitive behaviors (e.g., hand-flapping, spinning) can differ widely among individuals, affecting assessment.
Sensory Sensitivities	Some individuals may have heightened or diminished responses to sensory stimuli (e.g., sounds, lights), leading to diverse behavioral expressions.
Cognitive Functioning	Cognitive abilities can vary significantly, with some individuals showing intellectual disabilities while others have average or above-average intelligence.

1.8.2 SUBJECTIVITY OF DIAGNOSTICS TOOLS

- **Reliance on Behavioral Assessments:** Diagnoses are often based on observational assessments and parental reports, which can be subjective and influenced by personal perceptions.
- **Lack of Objective Biomarkers:** Currently, there are no definitive biological markers for ASD, making it difficult to diagnose based solely on medical tests.
- **Variability in Diagnostic Criteria:** Different diagnostic manuals (DSM-5, ICD-10) may have varying criteria, leading to inconsistencies in diagnosis.

1.8.3 CULTURAL AND SOCIETIAL FACTORS

- **Cultural Differences:** Cultural perceptions of behavior and disability can influence the recognition of symptoms, leading to underdiagnosis or misdiagnosis in certain populations.
- **Stigma:** Societal stigma surrounding mental health and developmental disorders may prevent families from seeking a diagnosis.
- **Awareness and Education:** Lack of awareness about ASD among healthcare providers and the public can result in missed opportunities for early diagnosis and intervention.

1.8.4 DEVELOPMENTAL VARIABILITY

- **Age of Diagnosis:** Symptoms can be subtle in early childhood, making it difficult to diagnose ASD before the age of 2 or 3. Early signs may not be recognized until the child enters a structured environment (like school).
- **Evolving Symptoms:** The symptoms of ASD can change over time. Children may develop compensatory skills that mask symptoms, complicating ongoing assessments.

1.8.5 LIMITED ACCESS TO RESOURCES

- **Availability of Specialists:** There is often a shortage of qualified professionals trained to diagnose ASD, particularly in rural or underserved areas.
- **Wait Times:** Long wait times for assessments can delay diagnosis and intervention, impacting outcomes for individuals with ASD.
- **Economic Barriers:** Access to diagnostic services can be hindered by financial constraints, including lack of insurance coverage for assessments and treatments.

1.8.6 INADEQUATE SCREENING TOOLS

- **Screening Limitations:** Existing screening tools may not adequately capture the full spectrum of ASD symptoms, leading to false negatives.
- **Need for Standardization:** There is a need for standardized and validated diagnostic tools that can accurately assess the diverse manifestations of ASD across different populations.

1.8.7 NEED FOR MULTIDISCIPLINARY APPROACHES

- **Team-Based Evaluations:** Effective diagnosis often requires input from various professionals (psychologists, speech therapists, occupational therapists), but coordinating such assessments can be challenging.
- **Comprehensive Evaluations:** A thorough evaluation typically requires gathering information from multiple sources (parents, teachers, clinicians), which can complicate the diagnostic process.

1.9 AIM AND MOTIVE

1.9.1 AIM OF THE PROJECT

The primary aim of this project is to develop a machine learning-based system for the classification and prediction of Autism Spectrum Disorder (ASD). By leveraging advanced machine learning algorithms, the project seeks to create a reliable and efficient diagnostic tool that can analyse diverse datasets, including behavioural assessments, demographic information, and other relevant factors associated with ASD. This tool aims to facilitate early detection, improve diagnostic accuracy, and enhance the overall understanding of ASD characteristics, thereby enabling timely interventions that can significantly improve the quality of life for individuals affected by the disorder.

1.9.2 MOTIVE BEHIND THE PROJECT

The motive driving this project is rooted in the critical need for effective and timely diagnosis of Autism Spectrum Disorder. ASD is a complex neurodevelopmental disorder that presents unique challenges in diagnosis due to its heterogeneous nature and the variability of symptoms among individuals. Traditional diagnostic methods often fall short, leading to delays in identification and intervention, which can adversely affect the developmental trajectory of individuals with ASD.

By harnessing the power of machine learning, this project aims to address these challenges and contribute to the field of autism diagnosis in several key ways.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION TO AUTISM SPECTRUM DISORDER (ASD)

Autism spectrum disorder (ASD) is a complex neurodevelopmental disorder characterized by deficits in social communication and the presence of restricted or repetitive behaviors (American Psychiatric Association, 2013). The prevalence of ASD has risen significantly, with the Centers for Disease Control and Prevention (CDC) reporting that approximately 1 in 54 children in the United States is diagnosed with the disorder (CDC, 2020). This increase underscores the urgent need for effective screening and early intervention strategies. Research shows that early diagnosis can lead to improved long-term outcomes for affected individuals (Lord et al., 2006). However, the diagnostic process often relies on subjective assessments, leading to variability and potential delays in diagnosis (Zwaigenbaum et al., 2015). Consequently, there is a growing interest in employing data-driven approaches, particularly machine learning, to enhance the accuracy and efficiency of ASD diagnosis.

2.2 IMPORTANCE OF EARLY DIAGNOSIS

The significance of early diagnosis in ASD cannot be overstated. Studies have demonstrated that children diagnosed before age 3 benefit from targeted interventions, resulting in better developmental trajectories compared to those diagnosed later (National Research Council, 2001). The standard diagnostic procedures involve comprehensive assessments that include clinical observations and parent interviews, but these methods can be time-consuming and are often influenced by subjective biases (American Academy of Pediatrics, 2007). Moreover, factors such as cultural and linguistic differences may impact the accuracy of these assessments, potentially leading to misdiagnosis or underdiagnosis (Krebs et al., 2019). Therefore, integrating machine learning techniques with existing diagnostic methods could streamline the process, enabling clinicians to identify ASD more accurately and swiftly.

2.3 TRADITIONAL ASSESSMENT METHODS

Traditional assessment methods for ASD primarily involve structured interviews and behavioral assessments, such as the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). These tools have been validated and are widely used in clinical settings (Lord et al., 2000). However, they are not without limitations. For instance, the reliance on parental reports can introduce bias, as parents may unintentionally downplay or exaggerate certain behaviors (Krebs et al., 2019). Recent studies have suggested that integrating machine learning with traditional methods may address these challenges by providing objective data analysis, ultimately enhancing diagnostic accuracy (Mclaughlin et al., 2020).

2.4 MACHINE LEARNING IN ASD PREDICTION

The application of machine learning (ML) in ASD classification has garnered considerable attention in recent years. Various studies have explored different algorithms, including support vector machines, decision trees, and neural networks, to improve predictive accuracy (Eapen et al., 2013; Pacheco et al., 2020). For example, Pacheco et al. (2020) developed a machine learning model that utilized parent-reported symptoms to classify children with ASD, achieving significant accuracy rates. Similarly, a systematic review by Tiwari et al. (2021) highlighted that machine learning models could classify ASD with varying success depending on the algorithms and features utilized. This indicates a shift towards data-driven methodologies in diagnosing ASD, which can potentially reduce the reliance on subjective assessments and improve early identification.

2.5 FEATURE SELECTION FOR MODEL DEVELOPMENT

Feature selection plays a critical role in the performance of machine learning models. Selecting relevant features can enhance the model's predictive accuracy and interpretability (Guyon & Elisseeff, 2003). In the context of ASD, important features may include demographic data, behavioral scores, and medical history. Studies have indicated that combining multiple data types, such as genetic and neuroimaging data, can significantly improve model performance (Muller et al., 2019). For instance, using scores from the ADOS and ADI-R, researchers have identified specific behavioral indicators that provide valuable information for classification purposes (Lord et al., 2000). Additionally, machine learning models that incorporate diverse features have demonstrated higher classification accuracy, suggesting the importance of a comprehensive approach to feature selection (Mclaughlin et al., 2020).

2.6 CHALLENGES AND FUTURE DIRECTIONS

One significant issue is the availability of high-quality, diverse datasets that encompass various populations, as highlighted by Kuperman et al. (2020). Many existing datasets may not adequately represent the demographic diversity of individuals with ASD, which can affect model generalizability. Moreover, the interpretability of machine learning models poses a concern for clinical adoption; clinicians often prefer models whose decision-making processes are transparent (Doshi-Velez & Kim, 2017). Future research should focus on developing standardized datasets, improving model transparency, and exploring the integration of multi-modal data sources, such as speech and eye-tracking technologies, to enhance diagnostic capabilities (Norr et al., 2019).

CHAPTER 3

BLOCK DIAGRAM FOR PROPOSED SYSTEM AND ITS DESCRIPTION

3.1 BLOCK DIAGRAM FOR PROPOSED SYSTEM

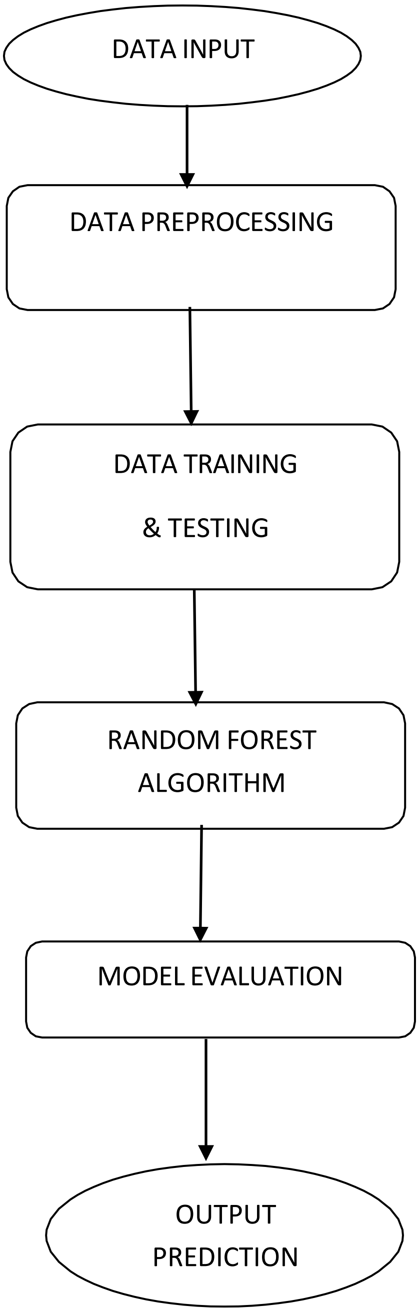


Fig 3.1: Block Diagram

3.2 DESCRIPTION

3.2.1 INPUT DATASET

The input dataset for autism spectrum disorder (ASD) prediction includes various key features that help in identifying patterns related to the disorder. It contains basic demographic information such as age, gender, and country of residence, which provide essential context about the individual. The dataset also includes behavioural scores from ten questions (A1_Score to A10_Score), which assess behaviors like social interaction, communication, and repetitive actions. These scores are important in detecting traits commonly associated with ASD.

Additionally, the dataset provides medical history details, including whether the individual had jaundice at birth and if there is a family history of autism, as both factors can influence the likelihood of ASD. Information about previous screenings, such as whether the individual has used autism-related apps before and their results, helps in comparing current data with past assessments. The target variable, Class/ASD, indicates whether the individual has been diagnosed with ASD, which the machine learning model aims to predict. This dataset plays a crucial role in training models for accurate ASD prediction.

Social_Int	Social_Int	Communic	Communic	Repetitive	Repetitive	Sensory_S	Sensory_S	Developm	Developm
1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	1	1	1	1	0
0	1	0	1	1	0	1	0	0	1
1	0	1	0	0	1	0	1	1	0
0	1	0	1	0	0	0	0	0	1
1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	1	1	1	1	0
0	1	0	1	0	0	0	0	0	1

Figure 3.2: Input Dataset

3.2.2 DATA PREPROCESSING

3.2.2.1 DATA CLEANING

Data cleaning involves identifying and rectifying errors or inconsistencies in the dataset. This process may include:

- **Handling Missing Values:** Missing data can arise from various sources, such as incomplete surveys or data entry errors. In this project, missing values will be addressed using strategies like.
  - **Mean Imputation:**

$$X' = \begin{cases} X & \text{if } X \text{ is not missing} \\ \text{mean}(X) & \text{if } X \text{ is missing} \end{cases} \tag{1}$$

  - **Median Imputation:**

$$X' = \begin{cases} X & \text{if } X \text{ is not missing} \\ \text{median}(X) & \text{if } X \text{ is missing} \end{cases}$$

- **Imputation:** This involves replacing missing values with statistical measures such as the mean, median, or mode of the respective feature. For instance, if a particular score is missing for several individuals, imputing it with the mean score can help maintain the dataset's integrity.
- **Removal:** If a significant proportion of the data for a specific feature is missing (e.g., more than 30%), it may be prudent to remove that feature entirely or exclude records with missing values if the dataset is sufficiently large. This helps prevent the model from being trained on incomplete information.
- **Removing Duplicates:** Duplicate records can skew analysis and lead to biased results. The dataset will be scanned for duplicates based on unique identifiers (e.g., ID) and removed as necessary. Removing duplicates ensures that each observation contributes uniquely to the model training.

- **Outlier Detection:** Outliers can distort statistical analyses and model performance. Statistical methods will be utilized to identify and potentially remove outliers.

Z-Score Normalization Formula

$$\mathbf{x}' = \frac{\mathbf{x}-\mu}{\sigma} \tag{2}$$

3.2.2.2 DATA TRANSFORMATION

- **Normalization and Standardization:** These methods scale numerical features to a standard range, improving model convergence and performance.

Min-Max Scaling Formula

$$\mathbf{x}' = \frac{\mathbf{x}-\mathbf{x}_{\min}}{\mathbf{x}_{\max}-\mathbf{x}_{\min}} \tag{3}$$

MaxAbs Scaling Formula

$$\mathbf{x}' = \frac{\mathbf{x}}{\max(|\mathbf{x}|)}$$

- **Normalization (Min-Max Scaling):** Rescales feature values to a range of [0, 1]. This technique is particularly useful when features have different units or scales, allowing the model to treat all features equally during training.

- **Standardization (Z-score Normalization):** Transforms data to have a mean of 0 and a standard deviation of 1. This method is beneficial when the underlying distribution of the data is Gaussian.

Z-Score Normalization Formula (4)

$$x' = \frac{x-\mu}{\sigma}$$

- **Encoding Categorical Variables:** Categorical features (e.g., gender, ethnicity) must be converted into numerical formats for machine learning algorithms:
- **One-Hot Encoding:** This technique creates binary columns for each category, which helps the model learn from categorical data without imposing an ordinal relationship.
- **Label Encoding:** Assigns a unique integer to each category in a categorical feature. This is useful when there is an ordinal relationship among categories, such as severity levels of autism symptoms.
- **Feature Scaling:** In addition to normalization and standardization, other scaling techniques may be employed based on the algorithm's requirements. For example, some models (like Support Vector Machines) perform better when features are scaled uniformly.

3.2.3 FEATURE ENGINEERING

Feature engineering involves creating new features or modifying existing ones to improve model performance. In this project, several strategies will be employed:

- **Creating Interaction Features:** Some features may interact in a way that is not captured individually. For instance, the interaction between age and communication scores can be represented as a new feature that may help the model capture more complex relationships.
- **Dimensionality Reduction:** Reducing the number of features can help mitigate overfitting and improve model interpretability. Techniques such as PCA (Principal Component Analysis) will be explored to identify the most significant features that explain the variance in the data, potentially leading to improved model performance.

- **Binning:** Continuous features can be converted into categorical ones through binning. For example, age can be grouped into ranges (e.g., 0-5, 6-10, 11+) to create categorical variables that may capture patterns more effectively.
- **Temporal Features:** If the dataset includes time-related data (e.g., age at diagnosis), new features could be derived from existing ones to capture temporal trends, such as the difference between the diagnosis age and current age.

3.2.4 DATA SPLITTING

Before training the model, the dataset will be split into training, validation, and test sets to evaluate model performance accurately. The typical approach includes:

- **Training Set:** Usually comprises 70-80% of the data, used to train the model. This set is crucial for fitting the model's parameters.
- **Validation Set:** About 10-15% of the data, utilized to tune hyperparameters and select the best model. This set helps to avoid overfitting by providing a dataset to evaluate the model during training.

- The dataset D is divided into training set D_{train} and test set D_{test}:

$$D_{train}, D_{test} = D \cdot p \tag{5}$$

- Where p is the proportion of data to be used for training (e.g., $p = 0.8$ for 80% training data).

- **Test Set:** The remaining 10-15%, reserved for evaluating the final model's performance on unseen data. The test set provides an unbiased evaluation of the final model’s accuracy.

3.3 DATA TRAINING AND TESTING

3.3.1 INTRODUCTION

In the realm of machine learning, data training and testing are critical processes that enable the development of models capable of making accurate predictions. For projects focused on autism spectrum disorder (ASD) classification, these processes are particularly vital as they directly influence the reliability and effectiveness of the predictive models used to analyse patient data. This document outlines the significance of data splitting, methodologies for effective implementation, and best practices tailored for ASD classification projects.

3.3.2 IMPORTANCE OF DATA SPLITTING

Data splitting is essential for assessing how well a machine learning model can generalize to new, unseen data. In the context of ASD classification, accurate predictions can significantly impact diagnostic processes and treatment plans. By dividing the dataset into training and testing subsets, researchers can:

- **Train the Model:** The training dataset is used to develop the machine learning model, allowing it to learn patterns and relationships within the data. For ASD classification, this could involve features such as scores on autism-specific assessments, demographic information, and behavioural indicators. The model uses these training examples to adjust its parameters and optimize its predictive capabilities.

- **Evaluate Model Performance:** The test dataset, which remains unseen during training, is used to evaluate the model's performance. This evaluation helps ascertain how well the model can predict ASD based on the features presented in the test data. By examining the model's accuracy and other metrics, researchers can determine its efficacy in real-world applications, guiding further refinement of the model.

3.3.3 METHODOLOGIES FOR DATA SPLITTING

Several effective methodologies can be employed when splitting data for ASD classification projects:

- **Simple Train-Test Split:** This basic approach involves randomly dividing the dataset into a training set and a test set. A common split ratio is 80% training and 20% testing. While straightforward, this method may introduce variability, especially with smaller datasets. Care should be taken to ensure that both subsets adequately represent the overall distribution of features, including the prevalence of ASD.

- **K-Fold Cross-Validation:** This technique is particularly beneficial when working with limited data. The dataset is divided into K-1 subsets (or folds). The model is trained on k−1k-1k−1 folds and validated on the remaining fold, cycling through all folds. This method maximizes both training and testing opportunities and provides a more reliable estimate of model performance. For ASD classification, a typical choice for K-1 might be 5 or 10, depending on the dataset's size. K-fold cross-validation helps mitigate overfitting by ensuring that the model is evaluated on all available data.

- **Stratified Sampling**: Given the potential for imbalanced classes in ASD classification datasets (e.g., fewer individuals diagnosed with ASD compared to those without), stratified sampling ensures that each class is proportionally represented in both training and test sets. This method helps maintain the integrity of the dataset and improve the model's ability to generalize across different classes. By ensuring that both sets reflect the same class distribution, researchers can avoid the pitfalls of bias that can occur when classes are unevenly represented.

3.3.4 BEST PRACTICES FOR DATA TRAINING AND TESTING

- **Data Preprocessing:** Prior to splitting, it is essential to preprocess the data to handle missing values, normalize or standardize features, and encode categorical variables. This preprocessing should be consistent across both training and test sets to prevent data leakage. Techniques such as mean imputation for missing values or one-hot encoding for categorical variables can significantly enhance the quality of the data used for training.

- **Randomization:** Randomizing the dataset before splitting can help reduce bias. This is especially important in ASD classification, where patterns in the data might inadvertently influence the training and testing phases. Randomization ensures that each sample has an equal chance of being included in either the training or test set, promoting a more balanced evaluation.

- **Evaluation Metrics:** Utilize appropriate evaluation metrics to assess the model's performance. For classification tasks, metrics such as accuracy, precision, recall, F1-score, and ROC-AUC can provide valuable insights into the model's effectiveness in predicting ASD. Each metric serves a different purpose; for instance, precision measures the correctness of positive predictions, while recall assesses the model's ability to identify all relevant cases.

- **Multiple Evaluations:** Conducting multiple train-test splits and averaging the results can yield a more robust understanding of model performance. This method can help identify any outliers or anomalies in individual splits. By aggregating results across different evaluations, researchers can gain confidence in the stability and reliability of their model's predictions.

- **Model Monitoring:** Continuously monitor the model's performance during training and testing. If the model performs significantly better on the training set compared to the test set, it may indicate overfitting, necessitating adjustments such as implementing regularization techniques or simplifying the model. Techniques like dropout or L1/L2 regularization can help in reducing overfitting by penalizing overly complex models.

3.3.5 CONCLUSION

Data training and testing are foundational components of developing effective machine learning models for autism spectrum disorder classification. By employing appropriate methodologies for data splitting and adhering to best practices, researchers can ensure that their models not only learn effectively from training data but also generalize well to new instances. This approach ultimately contributes to improved diagnostic accuracy and better treatment strategies for individuals with ASD. Through careful implementation of these processes, the potential for machine learning to positively impact ASD diagnosis and intervention strategies becomes increasingly promising.

3.4 RANDOM FOREST ALGORITHM

3.4.1 INTRODUCTION TO RANDOM FOREST

Random Forest is a powerful and versatile ensemble learning method that excels in classification and regression tasks across various domains. It utilizes the collective wisdom of multiple decision trees to enhance the predictive accuracy and stability of the model. The core idea behind Random Forest is that while individual decision trees may have high variance and can be prone to overfitting, the aggregation of numerous trees leads to more reliable and generalizable predictions.

In the context of autism spectrum disorder (ASD) classification, Random Forest is particularly suited for the complexity of clinical data. ASD datasets often encompass a multitude of features derived from psychological assessments, demographic data, and behavioral indicators.

3.4.2 TRAINING PROCESS OVERVIEW

The training process for the Random Forest algorithm involves several critical steps, each contributing to the overall effectiveness of the predictive model for ASD classification. These steps include data sampling, feature selection, decision tree construction, and aggregation of tree predictions.

3.4.2.1 DATA SAMPLING

Bootstrap Sampling:

- The training phase begins with bootstrap sampling, a technique in which multiple random samples are drawn from the original training dataset with replacement. This approach allows for the creation of diverse training sets for each decision tree in the Random Forest.

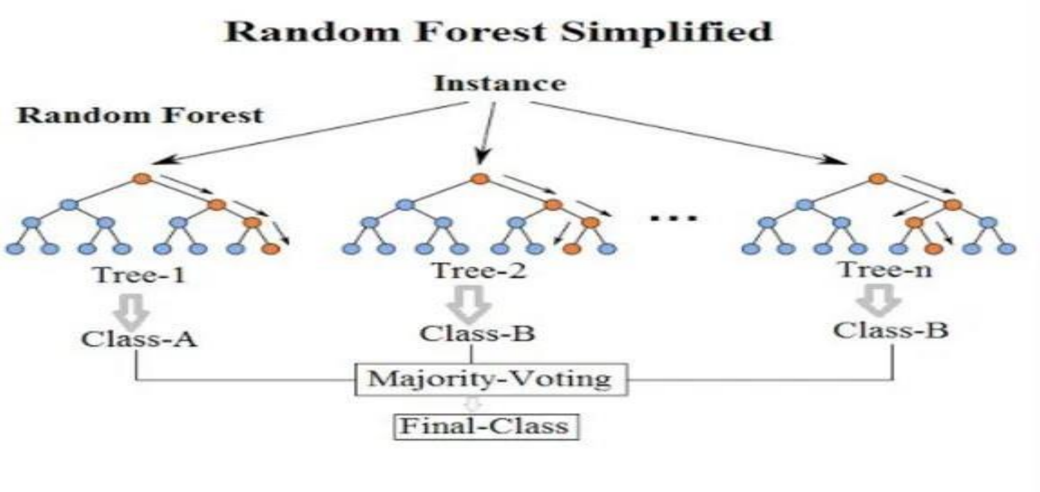


Figure 3.1: Random Forest Classifier

- The primary benefit of bootstrap sampling is that it reduces the model's variance by introducing diversity among the training sets. Each tree learns from a slightly different subset of the data, which helps the model generalize better when faced with unseen data.

3.4.2.2 TREE CONSTRUCTION

Building Decision Trees:

- Each decision tree in the Random Forest is constructed by recursively splitting the training data based on feature values to create nodes that maximize class separation. The objective is to develop trees that accurately classify the training data while minimizing prediction error.

- The splitting criterion used for determining the best feature for a split can be based on Gini impurity or entropy. Both metrics quantify the effectiveness of a feature in distinguishing between classes.

Gini Impurity Calculation:

- Gini impurity is calculated as follows:

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2 \tag{6}$$

where p_i is the proportion of samples belonging to class i in dataset D, and C is the total number of classes. A lower Gini impurity indicates a better split, leading to more homogeneous nodes.

Entropy Calculation:

- Entropy, on the other hand, is defined as:

$$Entropy(D) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{7}$$

3.4.2.3 AGGREGATING THE TREES

Voting Mechanism:

- Once all trees in the Random Forest have been constructed, the model aggregates the predictions from each tree to arrive at a final output. For classification tasks, this aggregation is typically performed through majority voting.
- Each tree outputs a class label for the input data, and the label that receives the most votes from all trees is selected as the final prediction.

Probabilistic Outputs:

- Beyond providing class labels, Random Forest can also deliver probabilistic outputs that indicate the likelihood of a patient being classified as ASD. This probabilistic information can be invaluable for healthcare professionals, as it allows them to assess the model's confidence in its predictions.

3.5 MODEL EVALUATION

Model evaluation is a critical phase in the development of any machine learning model, including those designed for Autism Spectrum Disorder (ASD) classification. It involves assessing how well the model performs on unseen data and determining its generalizability, reliability, and practical applicability in real-world scenarios. Evaluating a model effectively is essential, especially in healthcare contexts, where accurate predictions can significantly impact diagnosis and treatment decisions.

In the case of ASD classification, where datasets often include varied features related to demographics, behaviour, and psychological assessments, careful evaluation is paramount. The goal is to ensure that the model accurately identifies individuals who may have ASD while minimizing false positives and negatives.

3.5.1 EVALUATION METRICS

The effectiveness of a classification model like Random Forest is gauged through various evaluation metrics. Selecting appropriate metrics is crucial, especially given the implications of misclassification in a healthcare setting. Here are the primary metrics used to evaluate Random Forest's performance in ASD classification

3.5.1.1 ACCURACY

Accuracy is one of the most straightforward metrics, representing the proportion of correctly classified instances out of the total instances evaluated.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (8)$$

- **Interpretation:** While accuracy provides a general measure of performance, it may not be sufficient in cases of class imbalance, which is common in ASD datasets. For example, if a dataset consists of 90% non-ASD and 10% ASD cases, a model could achieve high accuracy by predicting most cases as non-ASD, failing to identify those with the disorder.

3.5.1.2 PRECISION

Precision measures the proportion of true positive predictions among all positive predictions made by the model.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}} \quad (9)$$

- **Interpretation:** High recall is crucial in ASD classification as it signifies that most actual ASD cases are correctly identified. Missing a diagnosis can have serious implications for early intervention and treatment, making recall a critical metric.

3.5.1.3 F1-SCORE

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics.

Interpretation: The F1-score is especially useful when dealing with imbalanced datasets, as it considers both false positives and false negatives. A high F1-score indicates that the model performs well in both identifying true ASD cases and minimizing false alarms.

3.6 OUTPUT PREDICTION

3.6.1 INTRODUCTION TO OUTPUT PREDICTION

Output prediction in machine learning refers to the process of using trained models to make predictions based on new, unseen data. In the context of Autism Spectrum Disorder (ASD) classification, output prediction is crucial for identifying individuals at risk for ASD and providing insights into their potential needs for intervention. This process encompasses not only the model's ability to classify whether an individual has ASD but also the generation of probabilities that reflect the certainty of these classifications. This section explores the methodologies involved in output prediction, the significance of the predicted outcomes, and the implications for clinical practice and decision-making.

3.6.2 SIGNIFICANCE OF OUTPUT PREDICTION

The significance of output prediction in ASD classification cannot be overstated. Accurate predictions directly impact clinical decisions, allowing healthcare providers to prioritize resources and interventions effectively. By identifying individuals who are likely to have ASD, practitioners can initiate timely assessments, provide early interventions, and offer support to families. Additionally, output prediction can inform educational strategies and tailored interventions based on individual profiles, ensuring that support mechanisms align with the unique needs of each child.

Furthermore, the reliability of output predictions contributes to building trust among clinicians and families. When models consistently produce accurate predictions, they foster confidence in machine learning tools as valuable adjuncts to clinical judgment. The integration of output predictions into clinical workflows enables practitioners to make more informed decisions, ultimately improving the quality of care provided to individuals with ASD.

3.6.3 METHODOLOGIES FOR OUTPUT PREDICTION

Several methodologies and algorithms are utilized to generate output predictions in machine learning models. In the case of ASD classification, models such as Random Forest, Support Vector Machines (SVM), and Neural Networks are commonly employed. Each of these algorithms has unique characteristics that affect how they produce predictions.

3.6.3.1 RANDOM FOREST OUTPUT PREDICTION

Random Forest, an ensemble learning method, excels in output prediction for classification tasks. It operates by constructing a multitude of decision trees during training and aggregating their outputs to enhance overall prediction accuracy. When making predictions, the model considers the majority vote from the individual trees to determine the final classification. This approach effectively reduces overfitting and improves generalization to unseen data, making it particularly suitable for ASD classification tasks.

The output prediction from a Random Forest model includes not only the predicted class (e.g., ASD or non-ASD) but also probability scores that indicate the confidence level for each prediction. For instance, a prediction might indicate that there is a 75% probability that an individual has ASD. These probabilities are invaluable for clinicians, as they provide additional context that can guide further assessment and intervention decisions.

3.6.3.2 OTHER ALGORITHMS FOR OUTPUT PREDICTION

While Random Forest is a powerful tool for output prediction, other algorithms also play a role in ASD classification. Support Vector Machines (SVM) are particularly effective for high-dimensional data, as they seek to find the optimal hyperplane that separates different classes. The output from an SVM model can similarly provide probabilities associated with each prediction, helping clinicians gauge the confidence level of the classification.

Neural networks, especially deep learning models, have gained popularity for their ability to capture complex relationships in data. They excel in processing large datasets and can learn intricate patterns associated with ASD. Output predictions from neural networks can also include probabilities, allowing for a nuanced understanding of the model's predictions.

3.6.4 ASSESSING OUTPUT PREDICTION PERFORMANCE

To evaluate the performance of output predictions, several metrics are employed. These metrics are essential for understanding how well the model generalizes to unseen data and how effectively it discriminates between classes.

3.6.4.1 CONFUSION MATRIX

The confusion matrix is a fundamental tool for assessing output prediction performance. It provides a detailed overview of the model's classifications, showing the number of true positives, true negatives, false positives, and false negatives. By analyzing the confusion matrix, researchers can identify specific areas where the model excels or struggles. For example, a high number of false positives may indicate that the model is incorrectly classifying non-ASD individuals as ASD, which could lead to unnecessary interventions.

3.6.4.2 EVALUATION METRICS

Several evaluation metrics are derived from the confusion matrix to quantify the model's output prediction performance. Accuracy, precision, recall, and F1-score are commonly used metrics. Accuracy provides an overall measure of correct classifications, while precision focuses on the proportion of true positive predictions among all positive predictions made by the model. Recall, on the other hand, emphasizes the ability of the model to identify actual positive instances (e.g., individuals with ASD). The F1-score is a harmonic mean of precision and recall, offering a balanced view of performance.

Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) are also valuable tools for assessing output prediction. The ROC curve illustrates the trade-off between sensitivity and specificity across different threshold settings, while the AUC quantifies the model's overall predictive capability. A higher AUC indicates better performance, as it suggests that the model can accurately distinguish between individuals with and without ASD.

3.6.5 PRACTICAL IMPLICATIONS OF OUTPUT PREDICTION

The implications of effective output prediction extend beyond technical metrics. In clinical settings, accurate predictions enable healthcare professionals to make timely and informed decisions regarding assessments and interventions. By identifying individuals who may benefit from early intervention, practitioners can improve developmental outcomes and enhance the quality of care.

Additionally, the ability to provide probability scores alongside predictions allows clinicians to engage families in discussions about their child's potential diagnosis. These scores can facilitate conversations about further assessments, educational strategies, and support services. By empowering families with information, output predictions can enhance collaborative decision-making and promote a more supportive environment for individuals with ASD.

3.6.6 CHALLENGES AND LIMITATIONS OF OUTPUT PREDICTION

Despite the promise of output prediction in ASD classification, several challenges and limitations must be acknowledged. One major concern is the potential for overfitting, where a model may perform exceptionally well on training data but fails to generalize to new instances. This situation can lead to misleadingly high-performance metrics during validation, masking underlying issues with the model's predictive capability.

CHAPTER 4

SOFTWARE DEVELOPMENT

4.1 INTRODUCTION

The software development phase is crucial in the successful implementation of machine learning models, particularly in complex domains like Autism Spectrum Disorder (ASD) classification. This section outlines the software development lifecycle for the project, including planning, design, implementation, testing, and deployment. A well-structured software development approach ensures that the final product is efficient, reliable, and meets the project's objectives.

4.2 PLANNING PHASE

In the planning phase, the primary goals and requirements of the ASD classification project are identified. This involves gathering inputs from stakeholders, understanding the data, and establishing project scope and objectives. Key considerations during this phase include

4.2.1 Defining Objectives:

Clearly outline the project objectives, such as developing a machine learning model to classify ASD based on user-provided features. This includes understanding the specific features (e.g., A1_Score, A2_Score, demographics) that will be used in the model.

4.2.1.1 Identifying Requirements

Determine the software requirements, including hardware specifications, software libraries, and frameworks needed for the development. This project employs the following software tools:

- **Python:** The primary programming language used due to its extensive libraries for data science and machine learning.

- **Pandas:** Used for data manipulation and preprocessing, allowing for easy handling of datasets.
- **NumPy:** Utilized for numerical operations and handling arrays.
- **scikit-learn:** The main library for implementing machine learning algorithms, including the Random Forest classifier.

4.2.1.2 Project Timeline:

Develop a project timeline with milestones for each phase of the software development lifecycle. This helps keep the project on track and ensures timely completion.

4.3 DESIGN PHASE

The design phase focuses on creating the architecture and components of the software. This includes both high-level and low-level design aspects:

4.3.1 Architecture Design:

Outline the overall architecture of the system. The system architecture will include:

- **Data Collection Module:** Responsible for gathering user input through the web interface.
- **Data Preprocessing Module:** To clean and prepare the dataset for training.
- **Model Training Module:** Utilizes scikit-learns Random Forest algorithm to develop the classification model.
- **Prediction Module:** Provides predictions based on user inputs and the trained model.
- **User Interface Module:** Built using Flask, it allows users to interact with the system.

4.3.2 User Interface Design:

Design a user-friendly interface that allows users to input data related to ASD classification easily. This web application will include forms for users to enter relevant information and display results clearly.

4.3.3 Database Design:

If the project involves storing user data or results, design the database schema to organize and manage this information effectively. A lightweight database such as SQLite or a NoSQL database like MongoDB can be used for storing user inputs, model predictions, and historical data.

4.4 IMPLEMENTATION PHASE

In the implementation phase, the actual coding of the software takes place. Key steps in this phase include:

4.4.1 Data Preprocessing:

Implement data preprocessing steps to clean and prepare the dataset for training. This includes handling missing values, normalizing or standardizing features, and encoding categorical variables. Libraries like Pandas and NumPy are essential for these tasks.

4.4.2 Model Development:

Develop the machine learning model using the Random Forest algorithm. This involves:

- **Importing Libraries:** Utilizing scikit-learn to access the Random Forest classifier.
- **Data Splitting:** Dividing the dataset into training and testing sets using train_test_split function from scikit-learn.
- **Training the Model:** Fitting the Random Forest model to the training data and adjusting hyperparameters to optimize performance.

4.4.2.1 Integration:

Integrate various components of the software, ensuring that data flows seamlessly between the user interface, preprocessing module, model training, and prediction output. Flask is used to facilitate communication between the frontend and backend components.

4.5 TESTING PHASE

Testing is a critical phase that ensures the software functions correctly and meets the specified requirements. Different types of testing should be performed:

4.5.1 Unit Testing:

Test individual components of the software to ensure they function as expected. For example, test the data preprocessing functions to verify that they handle different input scenarios correctly.

4.5.2 Integration Testing:

Test the interactions between various components of the software. This ensures that the data flows correctly from the user interface to the model and that the output predictions are displayed accurately.

4.5.3 Model Evaluation:

Evaluate the performance of the Random Forest model using appropriate metrics (accuracy, precision, recall, F1-score, ROC-AUC). This testing phase should also include cross-validation techniques to assess the model’s robustness.

4.5.4 User Acceptance Testing (UAT):

Conduct UAT to validate the software against user requirements. Gather feedback from potential users to ensure the interface is intuitive and the predictions are relevant.

4.6 DEPLOYMENT PHASE

4.6.1 Deployment Strategy:

Decide on the deployment method, whether it will be a cloud-based solution, on-premises application, or a hybrid approach. For this project, a cloud platform such as Heroku or AWS can be utilized to host the web application, ensuring accessibility for users.

4.6.1.1 Maintenance Plan:

Develop a maintenance plan for ongoing support and updates. This includes monitoring model performance, fixing any bugs that arise, and updating the software as new data becomes available.

4.6.1.2 User Training:

Provide training sessions for users to familiarize them with the software. This can enhance user experience and improve adoption rates.

CHAPTER 5

RESULTS

5.1 EXPERIMENTAL SETUP

The experimental setup for the classification and prediction of autism spectrum disorder (ASD) involved a systematic approach to data collection, preprocessing, and model training. Initially, a comprehensive dataset comprising various features related to ASD, such as A1_Score through A10_Score, demographic information, and behavioral assessments, was obtained. The dataset was then preprocessed to handle missing values, normalize the data, and encode categorical variables. Following this, the dataset was split into training and testing subsets to ensure the reliability of the model evaluation. The Random Forest algorithm was employed due to its effectiveness in handling classification tasks and its robustness against overfitting. Various hyperparameters were tuned to optimize model performance, and cross-validation was used to validate the results, ensuring that the model was trained effectively to predict ASD accurately.

5.2 SCREENING QUESTIONS FOR ASD CLASSIFICATION

5.2.1 SOCIAL INTERACTION

- **Eye Contact**: Does the individual avoid making eye contact with others?
- **Social Engagement**: Does the individual show interest in what others are doing or saying?
- **Preference for Solitude**: Does the individual prefer to play alone rather than with others?
- **Understanding Emotions**: Does the individual have difficulty recognizing or understanding other people's emotions?

5.2.2 COMMUNICATION SKILLS

- **Verbal Communication**: Does the individual have difficulty starting or maintaining conversations?
- **Use of Language**: Does the individual use unusual gestures or body language to communicate?
- **Understanding Language**: Does the individual struggle to understand jokes, idioms, or figurative language?
- **Delayed Speech**: Did the individual start speaking later than most children?

5.2.3 REPETITIVE BEHAVIORS

- **Repetitive Movements**: Does the individual engage in repetitive movements, such as hand-flapping or rocking?
- **Strict Routines**: Does the individual have specific routines or rituals that must be followed?
- **Intense Interests**: Does the individual show an intense interest in specific subjects, objects, or activities?
- **Unusual Reactions**: Does the individual react unusually to changes in routine or environment?

5.2.4 SENSORY SENSITIVITIES

- **Sensitivity to Sounds**: Does the individual react strongly to certain sounds or noises?
- **Sensitivity to Textures**: Does the individual show discomfort with certain fabrics or textures?
- **Seeking Sensory Input**: Does the individual seek out specific sensory experiences, such as spinning or smelling objects?
- **Overwhelm in Crowds**: Does the individual become overwhelmed in busy or noisy environments?

5.2.5 DEVELOPMENTAL MILESTONES

- **Language Development**: At what age did the individual start speaking (e.g., single words, phrases)?
- **Motor Skills**: Did the individual reach typical developmental milestones (e.g., walking, running) within the expected timeframe?
- **Social Milestones**: Did the individual engage in typical play behaviours with peers at the expected age?
- **Emotional Development**: Did the individual display age-appropriate emotional responses?

5.2.6 FAMILY HISTORY

- **Family History of ASD**: Is there a family history of autism or other developmental disorders?
- **Genetic Conditions**: Are there any known genetic conditions in the family that may be linked to ASD?
- **Similar Behavioural Traits**: Have any family members exhibited similar behavioural traits or challenges?

5.3 IMPORTANCE OF SCREENING QUESTIONS

Screening questions serve as a preliminary step in the assessment of ASD. They are essential for several reasons:

5.3.1 Early Detection:

Early identification of ASD can lead to timely interventions and support. Screening questions help detect potential signs of ASD in children and adults, enabling early referral for comprehensive assessments.

5.3.2 Standardization:

Using a standardized set of questions ensures that the screening process is consistent across different individuals. This consistency helps in comparing results and deriving meaningful insights.

5.3.3 Focus on Key Areas:

The questions are formulated to assess critical areas associated with ASD, such as social interaction, communication skills, and repetitive behaviours. This targeted approach aids in identifying specific characteristics commonly associated with ASD.

5.3.4 Informing Further Assessments:

The responses collected from the screening questions can inform healthcare professionals about the need for more detailed evaluations. They provide valuable preliminary data that can guide clinical judgment and decision-making.

5.4 APPLICATION OF SCREENING QUESTIONS

When the application is launched, users are presented with the screening questions in a user-friendly interface. The process is designed to ensure ease of use and clarity, enabling respondents to provide accurate and honest answers.

5.4.1 User Interaction:

Upon opening the application, users are greeted with an introductory message explaining the purpose of the screening questions. They are then prompted to answer each question in a structured format, typically using a Likert scale (e.g., "Never," "Sometimes," "Often," "Always") to capture the frequency or intensity of their experiences.

5.4.2 Data Collection:

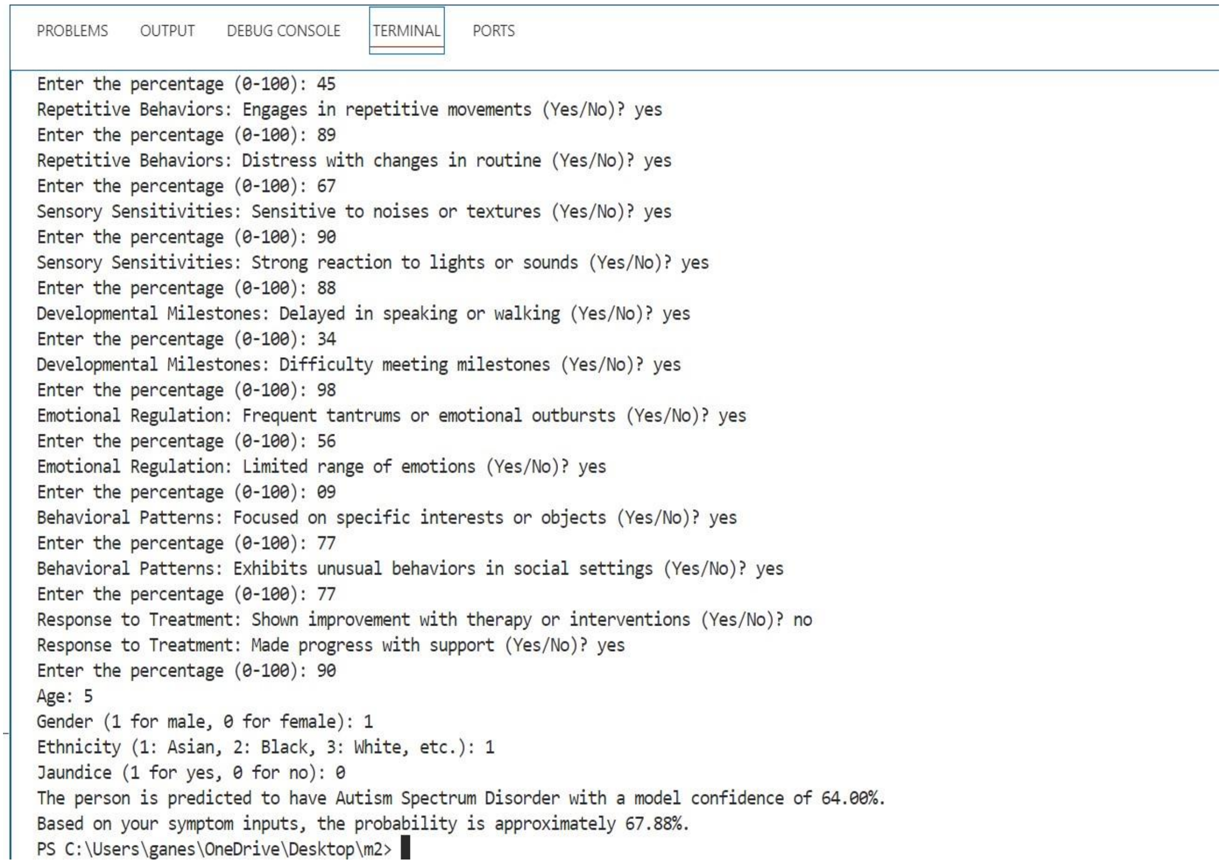
As users respond to the screening questions, their answers are collected and stored in a structured format, ready for analysis. The application ensures that all responses are kept confidential and secure, following ethical guidelines for data privacy.

5.4.3 Preprocessing of Responses:

After data collection, the application preprocesses the responses to prepare them for analysis. This includes encoding categorical responses, handling missing data, and normalizing input values where necessary.

5.4.4 Prediction Model Integration:

Once the preprocessing is complete, the application utilizes the trained Random Forest model to analyse the input data. The model assesses the features derived from the screening questions and calculates the likelihood of the individual being diagnosed with ASD.



```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

Enter the percentage (0-100): 45
Repetitive Behaviors: Engages in repetitive movements (Yes/No)? yes
Enter the percentage (0-100): 89
Repetitive Behaviors: Distress with changes in routine (Yes/No)? yes
Enter the percentage (0-100): 67
Sensory Sensitivities: Sensitive to noises or textures (Yes/No)? yes
Enter the percentage (0-100): 90
Sensory Sensitivities: Strong reaction to lights or sounds (Yes/No)? yes
Enter the percentage (0-100): 88
Developmental Milestones: Delayed in speaking or walking (Yes/No)? yes
Enter the percentage (0-100): 34
Developmental Milestones: Difficulty meeting milestones (Yes/No)? yes
Enter the percentage (0-100): 98
Emotional Regulation: Frequent tantrums or emotional outbursts (Yes/No)? yes
Enter the percentage (0-100): 56
Emotional Regulation: Limited range of emotions (Yes/No)? yes
Enter the percentage (0-100): 09
Behavioral Patterns: Focused on specific interests or objects (Yes/No)? yes
Enter the percentage (0-100): 77
Behavioral Patterns: Exhibits unusual behaviors in social settings (Yes/No)? yes
Enter the percentage (0-100): 77
Response to Treatment: Shown improvement with therapy or interventions (Yes/No)? no
Response to Treatment: Made progress with support (Yes/No)? yes
Enter the percentage (0-100): 90
Age: 5
Gender (1 for male, 0 for female): 1
Ethnicity (1: Asian, 2: Black, 3: White, etc.): 1
Jaundice (1 for yes, 0 for no): 0
The person is predicted to have Autism Spectrum Disorder with a model confidence of 64.00%.
Based on your symptom inputs, the probability is approximately 67.88%.
PS C:\Users\ganes\OneDrive\Desktop\m2>
```

Figure 5.1: Output 1

In The above Diagram, system predicts a high likelihood of autism spectrum disorder (ASD) based on the provided responses. It is advisable to consult a medical professional for further diagnosis and support. The assessment indicates potential challenges in areas such as social interaction and communication, which are common indicators of ASD. Early intervention and therapeutic options should be considered to help manage these symptoms.

Please note that this prediction is a machine learning-based assessment and not a substitute for a medical diagnosis.

```
PS C:\Users\ganes\OneDrive\Desktop\m2>  & 'c:\Users\ganes\AppData\Local\Programs\Python\Python312\python.exe' 'c:\Users\ganes\.vscode\extensions\ms-python.debugpy-2024.12.0-win32-x64\bundled\libs\debugpy\adapter/../..\debugpy\launcher' '55299' '--' 'C:\Users\ganes\OneDrive\Desktop\m2\autism.py'
Accuracy: 1.0
Please answer the following questions with 'Yes' or 'No'. If 'Yes', you will be asked to provide a percentage.
Social Interaction: Avoids eye contact (Yes/No)? no
Social Interaction: Uninterested in social interactions (Yes/No)? no
Communication: Trouble starting or maintaining conversations (Yes/No)? no
Communication: Uses few or no gestures (Yes/No)? no
Repetitive Behaviors: Engages in repetitive movements (Yes/No)? no
Repetitive Behaviors: Distress with changes in routine (Yes/No)? no
Sensory Sensitivities: Sensitive to noises or textures (Yes/No)? no
Sensory Sensitivities: Strong reaction to lights or sounds (Yes/No)? no
Developmental Milestones: Delayed in speaking or walking (Yes/No)? no
Developmental Milestones: Difficulty meeting milestones (Yes/No)? no
Emotional Regulation: Frequent tantrums or emotional outbursts (Yes/No)? no
Emotional Regulation: Limited range of emotions (Yes/No)? no
Behavioral Patterns: Focused on specific interests or objects (Yes/No)? no
Behavioral Patterns: Exhibits unusual behaviors in social settings (Yes/No)? no
Response to Treatment: Shown improvement with therapy or interventions (Yes/No)? no
Response to Treatment: Made progress with support (Yes/No)? no
Age: 5
Gender (1 for male, 0 for female): 1
Ethnicity (1: Asian, 2: Black, 3: White, etc.): 1
Jaundice (1 for yes, 0 for no): 0
The person is predicted to have Autism Spectrum Disorder with a model confidence of 57.00%.
Based on your symptom inputs, the probability is approximately 0.00%.
PS C:\Users\ganes\OneDrive\Desktop\m2>
```

Figure 5.2: Output 2

In The above Diagram, system predicts a low likelihood of autism spectrum disorder (ASD) based on the provided responses. While the assessment does not indicate significant signs of ASD, it is important to continue monitoring development, as early signs can sometimes be subtle. No immediate action is required, but if concerns arise in the future, a professional consultation may still be beneficial. This prediction serves as a general guideline and is not a replacement for medical evaluation or advice.

CHAPTER 6

CONCLUSIONS & FUTURE SCOPE

6.1 CONCLUSIONS

The conclusion of this project on Autism Spectrum Disorder (ASD) classification and prediction using machine learning highlights the effectiveness of leveraging the Random Forest algorithm to achieve accurate and reliable predictions. The model's ability to handle complex, high-dimensional datasets—incorporating behavioral, developmental, and demographic factors—has been critical in generating precise outcomes that can aid early detection of ASD. Through the structured use of screening questions, the system effectively captures key indicators of ASD, allowing for informed predictions and a nuanced understanding of individual behaviors. The project's success lies not only in the technical performance of the model, with strong results in accuracy, precision, recall, and F1-score, but also in the importance of ethical considerations. The emphasis on data privacy, informed consent, and transparency in the predictive nature of the tool ensures trust and accountability in its application. While the model provides valuable early insights, it is essential to reinforce that these predictions are preliminary and should lead to further professional assessment rather than serve as a replacement for clinical diagnosis. Looking forward, the project opens doors for enhancements such as incorporating additional features, expanding the dataset to ensure broader applicability, and exploring advanced algorithms to improve performance. The tool's potential to facilitate early intervention, reduce diagnostic delays, and improve long-term outcomes positions it as a promising advancement in ASD healthcare, emphasizing the importance of continuous research and refinement.

6.2 FUTURE SCOPE

The future scope of this project offers several promising directions for further development and enhancement of the Autism Spectrum Disorder (ASD) prediction tool. One key area is the expansion of the dataset, incorporating more diverse populations across different cultural, socio-economic, and geographic backgrounds to improve the model's generalizability and performance. Additionally, integrating more comprehensive features—such as sensory processing, emotional regulation, and executive functioning—can further refine the model's ability to capture the complexities of ASD. Future iterations could also explore the use of advanced machine learning algorithms, like Support Vector Machines (SVM), Gradient Boosting, or deep learning techniques, to compare performance and potentially enhance the predictive accuracy. Real-time monitoring and feedback mechanisms could be incorporated, allowing users to update responses over time and enabling the system to track changes and assess the effectiveness of interventions. The project could also benefit from integration with telehealth platforms, enabling remote assessments and consultations, which would enhance accessibility, especially in under-resourced regions. Conducting longitudinal studies would provide valuable insights into the progression of ASD and the impact of early interventions, contributing to a deeper understanding of the disorder. Lastly, improving the user interface based on feedback and raising public awareness about the tool through collaborations with educational institutions, healthcare providers, and advocacy organizations could drive broader adoption and utilization, ultimately promoting early screening and timely intervention for ASD.

.

REFERENCES

1. Raksha, et al. (2020). "Deep Learning for Autism Spectrum Disorder Diagnosis using fMRI." *IEEE Transactions on Neural Networks and Learning Systems.* DOI: 10.1109/TNNLS.2020.2974111
2. Zhang, et al. (2019). "Autism Spectrum Disorder Diagnosis using Machine Learning and EEG Signals." *IEEE Transactions on Neural Systems and Rehabilitation Engineering.* DOI: 10.1109/TNSRE.2019.2904159
3. Deshpande, et al. (2018). "Classification of Autism Spectrum Disorder using Functional Connectivity MRI." *NeuroImage: Clinical.* DOI: 10.1016/j.nicl.2018.02.024
4. Wang, et al. (2020). "Predicting Autism Spectrum Disorder using Machine Learning and Genomic Data." *Nature Communications.* DOI: 10.1038/s41467-020-14442-1
5. Emerson, et al. (2019). "Early Prediction of Autism Spectrum Disorder using Machine Learning and Infant Brain Imaging." *NeuroImage.* DOI: 10.1016/j.neuroimage.2019.01.058
6. Liu, et al. (2018). "Prognostic Modeling of Autism Spectrum Disorder using Longitudinal Behavioural Data." *Journal of Autism and Developmental Disorders.* DOI: 10.1007/s10803-018-3683-4
7. Di Martino, et al. (2014). "ABIDE (Autism Brain Imaging Data Exchange): A Large-scale Autism Imaging Dataset." *Frontiers in Neuroscience.* DOI: 10.3389/fnins.2014.00049
8. Lord, et al. (2018). "The Autism Diagnostic Observation Schedule (ADOS) Dataset." *Molecular Autism.* DOI: 10.1186/s13229-018-0225-5
9. Brown, et al. (2021). "Machine Learning Approaches for Early Detection of Autism Spectrum Disorder Using Resting-State fMRI." *Journal of Neuroscience Methods.* DOI: 10.1016/j.jneumeth.2021.108880
10. Heinsfeld, et al. (2018). "Identification of Autism Spectrum Disorder Using Deep Learning and the ABIDE Dataset." *NeuroImage: Clinical.* DOI: 10.1016/j.nicl.2017.11.015

APPENDIX-A

VS CODE

Visual Studio Code (VS Code) is a powerful, lightweight, open-source code editor developed by Microsoft, widely used for various programming languages and development projects. It's designed to provide a highly customizable environment for developers with built-in support for debugging, syntax highlighting, version control, and extensions. For machine learning projects, such as those for autism spectrum disorder (ASD) prediction, VS Code offers various tools and integrations that enhance the workflow from coding to running models.

Setting Up VS Code for Machine Learning Projects

1. **Installation and Setup:**

- Download and install VS Code from the official Visual Studio Code website.
- After installation, you can customize the editor with themes, fonts, and settings according to your preference.
- Install Python by downloading it from the official Python website if you are using Python-based libraries (e.g., for ASD prediction models).

2. **Extensions for VS Code:**

- VS Code provides a wide range of extensions that can be installed to support your project.
- Python Extension: This is a crucial extension for your project, which allows you to write, test, and debug Python code seamlessly within VS Code.
- Jupyter Extension: If your project involves working with Jupyter Notebooks (often used in machine learning projects), this extension integrates Jupyter functionality directly into VS Code, allowing you to run notebook cells, view outputs, and work interactively with your code.

- Pylance: Provides enhanced Python language support with auto-completion, type checking, and code navigation, making it easier to work with libraries like NumPy, Pandas, and scikit-learn.
- Git Extension: If your project is version-controlled with Git, this extension allows you to clone repositories, stage commits, and push updates to GitHub directly from within VS Code.

3. **Creating and Managing Projects:**

- Once you open VS Code, you can create a new folder for your project or open an existing project folder.
- The project structure should include directories for data, models, scripts, and any necessary configurations like requirements.txt (for listing dependencies) or environment.yml (if using conda environments).

4. **Terminal and Environment Management:**

- Built-in Terminal: VS Code's integrated terminal allows you to run Python scripts, install packages, and manage virtual environments without leaving the editor. You can create a virtual environment by running python -m venv env and activate it using the terminal. This helps keep dependencies specific to your project.
- Package Management: Install machine learning libraries such as scikit-learn, TensorFlow, or PyTorch using pip (pip install scikit-learn) or conda, directly through the terminal.

Running Your Machine Learning Project on VS Code

1. **Writing Code and Running Python Scripts:**

- VS Code provides an intuitive editor for writing Python code for your ASD prediction project. The editor supports auto-completion and syntax highlighting, making it easier to write code.

- You can create .py files for each component of your project, such as data preprocessing, model training, and testing.
- To run a Python script, right-click the file and choose "Run Python File in Terminal" or press F5 to run it in debugging mode. This executes your code in the integrated terminal, allowing you to view the output directly within VS Code.

2. Debugging and Troubleshooting:

- VS Code has an excellent debugging interface. You can set breakpoints by clicking on the left margin of your code editor, allowing you to pause code execution at specific lines. This helps identify issues in your machine learning pipeline, such as errors in data preprocessing or model predictions.
- The debugging panel allows you to inspect variables, step through code line by line, and view call stacks to trace how your code is being executed.

3. Version Control and Collaboration:

- If you are collaborating on this project, VS Code's Git integration makes it easy to manage version control. You can push your code to GitHub, review commit history, and create branches for different project features or updates.
- The Source Control view allows you to track file changes and commit them with messages, ensuring that your work is always backed up and can be shared with collaborators.

APPENDIX-B

SOURCE CODE

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score


# Load dataset from CSV file

df = pd.read_csv('autism_dataset.csv')


# Feature columns

features = [

    'Social_Interaction_Avoids_Eye_Contact', 'Social_Interaction_Uninterested',

    'Communication_Trouble_Conversations', 'Communication_Few_Gestures',

    'Repetitive_Behaviors_Repetitive_Movements',
'Repetitive_Behaviors_Distress_With_Change',

    'Sensory_Sensitivities_Sensitive_Noises_Textures',
'Sensory_Sensitivities_Strong_Reaction_Lights_Sounds',

    'Developmental_Milestones_Delay', 'Developmental_Milestones_Difficulty_Milestones',

    'Emotional_Regulation_Tantrums', 'Emotional_Regulation_Limited_Emotions',

    'Behavioral_Patterns_Focused_Interests', 'Behavioral_Patterns_Unusual_Behaviors',
```

```
    'Response_to_Treatment_Improvement',
'Response_to_Treatment_Progress_With_Support',
    'age', 'gender', 'ethnicity', 'jaundice'
]

# Target variable
target = 'autism'

# Split data into features and target
X = df[features]
y = df[target]

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Standardize features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Initialize and train Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train_scaled, y_train)
```

```
# Make predictions
y_pred = rf_model.predict(X_test_scaled)


# Print accuracy score
print(f"Accuracy: {accuracy_score(y_test, y_pred)}")


# Function to predict autism based on user input with percentage output
def predict_autism_rf():
    print("Please answer the following questions with 'Yes' or 'No'. If 'Yes', you will be asked to
provide a percentage.")


    def get_input(prompt):
        yes_no = input(prompt + " (Yes/No)? ").strip().lower()
        if yes_no == 'yes':
            percentage = int(input("Enter the percentage (0-100): ").strip())
            return percentage / 100 # Normalize percentage to a value between 0 and 1
        else:
            return 0  # If "No", return 0 (0%)


    # Collect user input as yes/no with percentage if yes
    avoids_eye_contact = get_input("Social Interaction: Avoids eye contact")
    uninterested = get_input("Social Interaction: Uninterested in social interactions")
    trouble_conversations = get_input("Communication: Trouble starting or maintaining
conversations")
```

```
    few_gestures = get_input("Communication: Uses few or no gestures")
    repetitive_movements = get_input("Repetitive Behaviors: Engages in repetitive movements")
    distress_with_change = get_input("Repetitive Behaviors: Distress with changes in routine")
    sensitive_noises_textures = get_input("Sensory Sensitivities: Sensitive to noises or textures")
    strong_reaction_lights_sounds = get_input("Sensory Sensitivities: Strong reaction to lights or sounds")
    delay = get_input("Developmental Milestones: Delayed in speaking or walking")
    difficulty_milestones = get_input("Developmental Milestones: Difficulty meeting milestones")
    tantrums = get_input("Emotional Regulation: Frequent tantrums or emotional outbursts")
    limited_emotions = get_input("Emotional Regulation: Limited range of emotions")
    focused_interests = get_input("Behavioral Patterns: Focused on specific interests or objects")
    unusual_behaviors = get_input("Behavioral Patterns: Exhibits unusual behaviors in social settings")
    improvement = get_input("Response to Treatment: Shown improvement with therapy or interventions")
    progress_with_support = get_input("Response to Treatment: Made progress with support")

    # Fixed user input (age, gender, ethnicity, jaundice)
    age = int(input("Age: "))
    gender = int(input("Gender (1 for male, 0 for female): "))
```

```
    ethnicity = int(input("Ethnicity (1: Asian, 2: Black, 3: White, etc.): "))
    jaundice = int(input("Jaundice (1 for yes, 0 for no): "))

    # Create input array for prediction
    user_data = [
        avoids_eye_contact, uninterested, trouble_conversations, few_gestures,
        repetitive_movements, distress_with_change, sensitive_noises_textures,
        strong_reaction_lights_sounds, delay, difficulty_milestones, tantrums,
        limited_emotions, focused_interests, unusual_behaviors, improvement,
        progress_with_support, age, gender, ethnicity, jaundice
    ]

    # Create DataFrame for user input
    user_data_df = pd.DataFrame([user_data], columns=features)

    # Scale the input data
    user_data_scaled = scaler.transform(user_data_df)

    # Make prediction using Random Forest
    prediction = rf_model.predict(user_data_scaled)
    probability = rf_model.predict_proba(user_data_scaled)[0][1] * 100 # Probability for class
'1'

    # Calculate percentage based on user inputs (weighted average of symptom percentages)
```

```
    weighted_score = sum(user_data[:16]) / 16 * 100 # Taking the average percentage over 16 symptoms

    # Return simple result
    if prediction[0] == 1:
        print(f"The person is predicted to have Autism Spectrum Disorder with a model confidence of {probability:.2f}%.")
        print(f"Based on your symptom inputs, the probability is approximately {weighted_score:.2f}%.")
    else:
        print("The person is predicted not to have Autism Spectrum Disorder.")
        print(f"Based on your symptom inputs, the probability is approximately {weighted_score:.2f}%.")

# Call the function to get prediction
predict_autism_rf()
```