A MACHINE LEARNING FRAMEWORK FOR EARLY-STAGE DETECTION OF AUTISM SPECTRUM DISORDERS

A PROJECT REPORT

Submitted by

KAUSHIK S [211420104125]
KRISHNARAGHAVAN M [211420104140]
MANIGANDAN A [211420104155]

in partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

APRIL 2024

PANIMALAR ENGINEERING COLLEGE
(An Autonomous Institution, Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this project report "**A MACHINE LEARNING FRAMEWORK FOR EARLY-STAGE DETECTION OF AUTISM SPECTRUM DISORDERS**" the bonafide work of KAUSHIK S [211420104125], KRISHNARAGHAVAN M [211420104140], MANIGANDAN A [211420104155] who carried out the project work under my supervision.

SIGNATURE

Dr.L.JABASHEELA,M.E.,Ph.D.,
HEAD OF THE DEPARTMENT
DEPARTMENT OF CSE,
PANIMALAR ENGINEERING COLLEGE,
NASARATHPETTAI,
POONAMALLEE,
CHENNAI-600 123.

SIGNATURE

Dr.L.JABASHEELA,M.E.,Ph.D.,
HEAD OF THE DEPARTMENT
DEPARTMENT OF CSE,
PANIMALAR ENGINEERING COLLEGE,
NASARATHPETTAI,
POONAMALLEE,
CHENNAI-600 123.

Certified that the above mentioned students were examined in End Semester Project Viva- Voce held on.....

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION BY THE STUDENT

We KAUSHIK S, KRISHNARAGHAVAN M, MANIGANDAN A (Reg.No) (211420104125), (211420104140), (211420104155) hereby declare that this project report titled "**A Machine learning framework for early-stage detection of Autism Spectrum Disorders**", under the guidance of **Dr.L.JABASHEELA ,M.E., Ph.D.,** is the original work done by us and we have not plagiarized or submitted to any other degree in any university by us.

KAUSHIK S

KRISHNARAGHAVAN M

MANIGANDAN A

ACKNOWLEDGEMENT

We would like to express our deep gratitude to our respected Secretary and Correspondent **Dr. P.CHINNADURAI, M.A., Ph.D.** for his kind words and enthusiastic motivation, which inspired us a lot in completing this project.

We express my sincere thanks to our beloved Directors **Tmt. C.VIJAYARAJESWARI**, **Dr. C.SAKTHI KUMAR, M.E., Ph.D** and **Dr. SARANYASREE SAKTHI KUMAR B.E., M.B.A., Ph.D.,** for providing us with the necessary facilities to undertake this project.

We also express our gratitude to our Principal **Dr. K.Mani, M.E., Ph.D.** who facilitated us in completing the project.

We thank the Head of the CSE Department, **Dr. L.JABASHEELA, M.E., Ph.D.,** for the support extended throughout the project.

We would like to express our sincere thanks to **DR. G. SENTHILKUMAR, M.E., Ph.D.,** and all the faculty members of the Department of CSE for their advice and encouragement for the successful completion of the project.

KAUSHIK S

KRISHNARAGHAVAN M

MANIGANDAN A

PROJECT COMPLETION CERTIFICATE

(ONE PAGE ONLY)

ABSTRACT

Autism Spectrum Disorder (ASD) presents significant social and communication challenges, affecting 1 in every 59 children. Early diagnosis is critical for timely intervention, yet many individuals are diagnosed later in life, hindering access to support. Our research employs logistic regression, a powerful computational tool, to enhance the accuracy and efficiency of ASD identification. Through systematic analysis, we identify behavioural markers associated with ASD, training the logistic regression model to distinguish between ASD and neurotypical individuals.

Functioning like a skilled investigator, the logistic regression model meticulously examines data to discern subtle patterns indicative of ASD. As the model iteratively learns, its ability to accurately identify ASD improves significantly. Our ultimate goal is to develop an accessible online tool for swiftly and accurately identifying ASD in children. Such a tool has the potential to revolutionize diagnostics, facilitating early intervention and support.

Moreover, implementing logistic regression in ASD diagnosis holds broader implications for reducing healthcare costs and improving outcomes. By streamlining diagnostics and enabling timely access to support, we aim to alleviate the burden of ASD and enhance the quality of life for affected individuals and families. Our research underscores the transformative potential of logistic regression in advancing early ASD detection, promising improved outcomes and enhanced support for those affected by the disorder.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNN	-	Convolutional Neural Network
LCNN	-	Lookup based Convolutional Neural Network
RNN	-	Recurrent Neural Network
DEX	-	Dalvik Executables
TCP	-	Transmission Control Protocol
IP	-	Internet Protocol
HTTP	-	Hyper Text Transfer Protocol
ADT	-	Android Development Tool

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	APPENDICES	
A.1	SDG Goals	41
A.2	Source Code	41
A.3	Screen Shots	47
A.4	Plagiarism Report	53
	REFERENCES	65

CHAPTER 1
INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1 Overview

Early diagnosis of Autism Spectrum Disorder (ASD) is crucial for better outcomes. This project focuses on developing a Machine Learning (ML) framework specifically tailored for the early detection of Autism Spectrum Disorder (ASD) in toddlers. By analyzing data exclusively from this age group, the framework aims to enhance early diagnosis, which is critical for improving outcomes for individuals with ASD.

The project will utilize various ML algorithms to analyze data collected from toddlers exhibiting potential ASD symptoms. By leveraging age-appropriate features and characteristics, the framework aims to identify the most effective ML methods for detecting ASD in this specific age range. Additionally, it will pinpoint the key factors that reliably predict ASD in toddlers, enabling more accurate and timely diagnosis.

This data-driven approach has the potential to significantly revolutionize ASD detection in toddlers. Unlike traditional methods, ML offers the promise of faster and more objective screening processes. Imagine a framework that recommends the most suitable ML approach based solely on a toddler's data, significantly enhancing the efficiency and accuracy of ASD diagnosis at an early stage. This, in turn, can facilitate earlier interventions, leading to improved outcomes for toddlers with ASD.

Despite its potential, the project acknowledges several challenges. High-quality data collection is paramount to train accurate ML models for toddler-specific ASD detection. Moreover, ensuring the interpretability of these models is crucial for establishing trust in their results. Ethical considerations, particularly regarding potential biases in the data and the responsible use of ML in healthcare, must also be addressed.

By overcoming these challenges and focusing specifically on toddlers' data, this project has the potential to be a transformative force in early ASD detection, ultimately improving the lives of toddlers and their families affected by ASD.

1.2 Problem Definition

ASD, a neurodevelopmental disorder impacting communication and social interaction, necessitates early detection for effective intervention and improved long-term outcomes. Current diagnostic methods, reliant on behavioural observations and standardized tests, are time-consuming, costly, and subjective. Hence, there's a need for:

- **Objective Methods:** Existing approaches may be influenced by clinician experience or a child's behaviour on assessment days.
- **Early Detection:** Identifying ASD earlier allows for timelier intervention initiation.
- **Enhanced Accessibility:** More accessible and affordable diagnostic tools could facilitate early identification, particularly in underserved communities.

Machine Learning (ML) offers promise in addressing these challenges. By analyzing various data sources (e.g., brain scans, eye tracking, speech patterns), ML models could:

- **Automate Diagnostic Processes:** Streamlining diagnosis and reducing costs.
- **Detect Unseen Patterns:** ML algorithms can uncover subtle ASD indicators hidden in vast datasets.
- **Create Screening Tools**: Developing quick and easy-to-administer ASD screening tools.

However, hurdles persist:

- **Data Quality and Quantity:** Large, high-quality datasets of toddler-specific ASD indicators are required for training accurate ML models.
- **Interpretability:** Understanding the reasoning behind ML model decisions is crucial for trust-building.
- **Ethical Considerations:** Addressing biases in data and guarding against potential ML model misuse in healthcare settings.

In summary, the problem definition involves crafting reliable and ethically sound ML-based tools for early ASD detection in toddlers. This aims to enhance diagnosis precision, intervention efficacy, and outcomes for individuals with ASD.

CHAPTER-2
LITERATURE SURVEY

CHAPTER-2
LITERATURE SURVEY

1.TITLE: Efficient machine learning models for early stage detection of autism spectrum disorder [1]

YEAR: 2022

AUTHORS: M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni

ABSTRACT:

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by persistent deficits in social communication and interaction, as well as restricted, repetitive patterns of behavior, interests, or activities. Early detection of ASD is crucial for timely intervention and improved outcomes. In this study, we propose efficient machine learning models for the early stage detection of ASD. We utilize a dataset comprising various behavioral features and demographic information collected from individuals diagnosed with ASD and typically developing individuals. Through rigorous experimentation and feature selection techniques, we identify the most discriminative features for ASD detection. Subsequently, we train and evaluate multiple machine learning models, including support vector machines, decision trees, random forests, and neural networks, to classify individuals as ASD or non-ASD. Our results demonstrate the effectiveness of the proposed models in achieving high accuracy, sensitivity, and specificity in early stage ASD detection. The developed models hold promise for assisting clinicians and healthcare professionals in identifying individuals at risk of ASD at an early age, facilitating timely intervention and support.

2.TITLE: Machine learning data analysis highlights the role of parasutterella and alloprevotella in autism spectrum disorders [2]

YEAR: 2022

AUTHORS: D. Pietrucci, A. Teofani, M. Milanesi, B. Fosso, L. Putignani, F. Messina, G. Pesole, A. Desideri, and G. Chillemi

ABSTRACT:

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by heterogeneous symptoms and etiologies. Recent research has suggested a potential link between gut microbiota composition and ASD. In this study, we utilize machine learning data analysis techniques to investigate the role of specific gut microbial taxa in ASD. By analyzing

metagenomic data from individuals with ASD and neurotypical controls, we identify differential abundances of Parasutterella and Alloprevotella species in the gut microbiota of individuals with ASD. Furthermore, our machine learning models reveal associations between the abundance of these microbial taxa and ASD symptom severity. These findings provide insights into the potential involvement of Parasutterella and Alloprevotella in the pathophysiology of ASD and highlight the utility of machine learning approaches in microbiome research. Further investigation is warranted to elucidate the mechanisms underlying gut-brain interactions in ASD and explore the therapeutic implications of targeting specific microbial taxa for ASD management.

3.TITLE: A fuzzy-based eye gaze point estimation approach to study the task behavior in autism spectrum disorder [4]

YEAR: 2018

AUTHORS: J. Amudha and H. Nandakumar

ABSTRACT:

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by difficulties in social communication and interaction, as well as restricted and repetitive patterns of behavior. Eye gaze behavior is a crucial aspect of social interaction and communication, and studying gaze patterns can provide insights into the cognitive and social processes underlying ASD. In this study, we propose a fuzzy-based approach for estimating eye gaze points to analyze task behavior in individuals with ASD. Our approach utilizes eye tracking technology to capture gaze data during task performance. We employ fuzzy logic techniques to process and analyze the gaze data, allowing for the estimation of gaze points with enhanced accuracy and robustness. By examining task behavior through the lens of eye gaze patterns, we aim to uncover subtle differences in cognitive processing between individuals with ASD and neurotypical individuals. The insights gained from this study have the potential to inform the development of personalized interventions tailored to the specific needs of individuals with ASD. By applying our approach to tasks designed to assess social cognition and attention in individuals with ASD, we aim to identify specific gaze patterns associated with ASD-related behaviors. The results of our study contribute to a better understanding of the cognitive and behavioral characteristics of individuals with ASD and may inform the development of more effective interventions and support strategies.

4.TITLE: Aarya_A kinesthetic companion for children with autism spectrum disorder [3]

YEAR: 2017

AUTHORS: R. Sreedasyam, A. Rao, N. Sachidanandan, N. Sampath, and S. K. Vasudevan

ABSTRACT:

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by challenges in social interaction, communication, and repetitive behaviors. Children with ASD often face difficulties in engaging with others and may benefit from interventions that promote social interaction and communication skills. In this paper, we present "Aarya," a kinesthetic companion designed to support children with ASD in their social and emotional development. Aarya utilizes interactive technology and kinesthetic feedback to engage children in various activities aimed at improving social communication, emotional regulation, and sensory integration skills. The system incorporates elements of play therapy and behavioral therapy, providing a safe and supportive environment for children to explore and interact. The companion also incorporates personalized feedback mechanisms to adapt to the individual needs and preferences of each child. In addition to promoting social interaction, Aarya offers opportunities for sensory stimulation and regulation, which are particularly beneficial for children with sensory processing difficulties commonly associated with ASD. The effectiveness of Aarya as a therapeutic tool is evaluated through pilot studies involving children diagnosed with ASD, their parents, and therapists. Preliminary results indicate positive outcomes in terms of increased engagement, social interaction, and emotional regulation among participants.Future research aims to further refine and validate the effectiveness of Aarya through larger-scale clinical trials and longitudinal studies.

5.TITLE: A deep learning approach to predict autism spectrum disorder using multisite resting-state fMRI [6]

YEAR: 2021

AUTHORS: F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba

ABSTRACT:

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by atypical brain connectivity patterns. Resting-state functional magnetic resonance imaging (rs-fMRI) has emerged as a valuable tool for investigating neural connectivity in individuals with ASD. In this study, we propose a deep learning approach for predicting ASD using multisite rs-fMRI data. Our method leverages convolutional neural networks (CNNs) to extract

features from rs-fMRI scans acquired from multiple sites, addressing challenges related to data heterogeneity and site-specific variability. We employ transfer learning techniques to fine-tune pre-trained CNN models on rs-fMRI data, enabling the classification of individuals as ASD or typically developing (TD). Through rigorous experimentation and cross-validation, we demonstrate the efficacy of our approach in accurately predicting ASD status across different sites. Furthermore, we conduct feature visualization analyses to elucidate the neural substrates underlying ASD classification. The proposed deep learning framework offers a promising avenue for non-invasive and objective ASD diagnosis based on neuroimaging biomarkers. Our findings contribute to advancing the understanding of ASD pathophysiology and may facilitate early intervention and personalized treatment strategies for individuals with ASD

6.TITLE: The contribution of machine learning and eye-tracking technology in autism spectrum disorder research [7]

YEAR: 2021

AUTHORS: K.-F. Kollias, C. K. Syriopoulou-Delli, P. Sarigiannidis, and G. F. Fragulis

ABSTRACT:

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental condition characterized by impairments in social communication and interaction, as well as restricted and repetitive behaviors. Over the years, various technological advancements have been explored to aid in the understanding and diagnosis of ASD. In this systematic review, we investigate the contribution of machine learning and eye-tracking technology in ASD research. We conduct a comprehensive review of studies published in peer-reviewed journals, focusing on the application of machine learning algorithms and eye-tracking techniques to study ASD-related behaviors and cognitive processes. Our review synthesizes findings from studies employing machine learning for ASD diagnosis, symptom severity prediction, and classification of ASD subtypes. Additionally, we examine the role of eye-tracking technology in elucidating gaze patterns, visual attention, and social cognition in individuals with ASD. Through the integration of machine learning and eye-tracking approaches, researchers have made significant strides in identifying objective biomarkers and quantifying behavioral phenotypes associated with ASD. However, challenges remain, including data heterogeneity, sample size limitations, and the need for standardized methodologies. Future research should aim to address these challenges and further explore the potential of machine learning and eye-tracking technology in advancing our understanding of ASD and improving diagnostic accuracy and intervention strategies.

7.TITLE: Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques [8]

YEAR: 2022

AUTHORS: I. A. Ahmed, E. M. Senan, T. H. Rassem, M. A. H. Ali, H. S. A. Shatnawi, S. M. Alwazer, and M. Alshahrani

ABSTRACT:

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by difficulties in social communication and interaction, as well as restricted and repetitive behaviors. Early diagnosis and intervention are crucial for improving outcomes for individuals with ASD. In this study, we propose an eye tracking-based approach for the diagnosis and early detection of ASD using machine learning and deep learning techniques. Our method leverages eye tracking technology to analyze gaze patterns and visual attention in individuals with ASD and neurotypical controls. We extract features from eye tracking data and employ machine learning and deep learning algorithms to develop predictive models for ASD diagnosis and early detection. Through rigorous experimentation and cross-validation, we demonstrate the efficacy of our approach in accurately identifying individuals with ASD and distinguishing them from neurotypical individuals. Furthermore, we investigate the contribution of different gaze metrics and regions of interest to ASD classification, providing insights into the underlying neural mechanisms associated with ASD-related behaviors. The proposed approach offers a non-invasive and objective method for ASD diagnosis and early intervention, facilitating timely support for affected individuals and their families.

8.TITLE: Identification and exploration of facial expression in children with ASD in a contactless environment [10]

YEAR: 2019

AUTHORS: S. P. Abirami, G. Kousalya, and R. Karthick

ABSTRACT:

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by challenges in social communication, interaction, and the presence of repetitive behaviors or restricted interests. One significant difficulty faced by individuals with ASD is in understanding and interpreting facial expressions, which are fundamental for effective social

communication. The ability to recognize emotions conveyed through facial expressions plays a crucial role in navigating social interactions and forming meaningful relationships. However, individuals with ASD often struggle with accurately identifying and interpreting facial cues, which can lead to misunderstandings, social isolation, and difficulties in building and maintaining relationships. The proposed approach holds significant promise for enhancing our understanding of the social and emotional difficulties faced by children with ASD. By automatically recognizing and analyzing facial expressions, researchers can gain valuable insights into the emotional experiences of children with ASD in various contexts. Additionally, the contactless nature of the imaging devices minimizes potential discomfort for children with ASD, facilitating their participation in research studies and clinical assessments. By leveraging computer vision techniques and machine learning algorithms, researchers can shed light on the social and emotional difficulties experienced by individuals with ASD, ultimately paving the way for improved support and outcomes for this population.

9.TITLE: Detecting autism spectrum disorder using machine learning techniques [11]

YEAR: 2021

AUTHORS: M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam

ABSTRACT:

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder characterized by challenges in social communication, interaction, and repetitive behaviors. Early detection and diagnosis of ASD are crucial for timely intervention and support. In this study, we investigate the use of machine learning techniques for the detection of ASD. Leveraging a variety of data sources, including clinical assessments, behavioral observations, and developmental history, we develop and evaluate machine learning models for ASD detection. Our approach aims to integrate multiple features and modalities to enhance the accuracy and robustness of ASD detection algorithms. Through rigorous experimentation and validation, we demonstrate the effectiveness of our machine learning approach in accurately identifying individuals with ASD. The proposed framework holds promise for aiding clinicians and healthcare professionals in the early detection and diagnosis of ASD, facilitating timely intervention and support for affected individuals and their families. By leveraging machine learning techniques, we contribute to the development of objective and data-driven approaches for ASD detection,

addressing the challenges associated with traditional diagnostic methods and improving outcomes for individuals with ASD.

10. TITLE: A new computational intelligence approach to detect autistic features for autism screening [13]

YEAR: 2018

AUTHORS: F. Thabtah, F. Kamalov, and K. Rajab

ABSTRACT:

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder characterized by impairments in social communication and interaction, as well as restricted and repetitive behaviors. Early detection and intervention are essential for improving outcomes for individuals with ASD. Thabtah, Kamalov, and Rajab introduce a novel computational intelligence approach for detecting autistic features to aid in autism screening. Leveraging machine learning and computational intelligence techniques, the proposed approach analyzes various behavioral and clinical data to identify patterns indicative of ASD. By integrating multiple data sources and employing advanced data analysis methods, the authors aim to develop a robust screening tool capable of accurately identifying individuals with ASD. Through comprehensive evaluation and validation, the effectiveness of the computational intelligence approach in detecting autistic features is demonstrated. This research contributes to the development of efficient and reliable screening methods for ASD, facilitating early intervention and support for affected individuals and their families. Additionally, the adoption of computational intelligence techniques offers a promising avenue for enhancing the accuracy and accessibility of ASD screening, ultimately improving outcomes and quality of life for individuals with ASD. The proposed approach holds potential for integration into existing healthcare systems, providing clinicians and healthcare professionals with a valuable tool for early identification and intervention in ASD.

11. TITLE: Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison [14]

YEAR: 2021

AUTHORS: M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni

ABSTRACT:

Heart disease is a leading cause of mortality worldwide, emphasizing the importance of accurate prediction and early intervention. In this study, Ali et al. explore the use of supervised

machine learning algorithms for heart disease prediction, aiming to analyze performance and compare different models. Leveraging a diverse dataset containing clinical and demographic features, the authors employ various machine learning algorithms to develop predictive models for heart disease. Through extensive performance analysis and comparison, including evaluation metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC), the study provides insights into the effectiveness of different algorithms in predicting heart disease risk. The results highlight the potential of machine learning approaches in enhancing the accuracy and efficiency of heart disease prediction, facilitating timely intervention and personalized healthcare. By leveraging advanced data analysis techniques, clinicians and healthcare professionals can improve risk assessment and management strategies, ultimately reducing the burden of heart disease and improving patient outcomes. This research underscores the significance of data-driven approaches in advancing cardiovascular medicine and highlights the potential for machine learning to revolutionize cardiac risk prediction and management.

12. TITLE: Stroke risk prediction with machine learning techniques [15]

YEAR: 2022

AUTHORS: E. Dritsas and M. Trigka

ABSTRACT:

Stroke is a significant cause of mortality and disability worldwide, highlighting the importance of accurate risk prediction and prevention strategies. In this study, Dritsas and Trigka investigate the use of machine learning techniques for stroke risk prediction, aiming to develop accurate and reliable predictive models. Leveraging a comprehensive dataset containing clinical, demographic, and lifestyle factors, the authors employ various machine learning algorithms to assess stroke risk. Through rigorous analysis and evaluation, including performance metrics such as sensitivity, specificity, and area under the receiver operating characteristic curve (AUC), the study evaluates the effectiveness of different machine learning techniques in predicting stroke risk. The results demonstrate the potential of machine learning approaches in enhancing stroke risk assessment, enabling early intervention and personalized prevention strategies. By leveraging advanced data analysis techniques, clinicians and healthcare professionals can improve risk prediction models, ultimately reducing the burden of stroke-related morbidity and mortality. This research underscores the importance of data-driven approaches in stroke prevention and highlights the role of machine learning in advancing

predictive analytics for stroke risk assessment. The findings of this study contribute to the growing body of literature aimed at enhancing stroke prevention efforts through the application of machine learning techniques, paving the way for improved patient outcomes and reduced healthcare costs.

CHAPTER-3
THEORETICAL BACKGROUND

CHAPTER-3
THEORETICAL BACKGROUND

3.1 EXISTING SYSTEM

This research aims to create an effective prediction model using different types of ML methods to detect autism in people of different ages. First of all, the datasets are collected, and then the preprocessing is accomplished via the missing values imputation, feature encoding, and oversampling. The Mean Value Imputation (MVI) method is used to impute the missing values of the dataset. Then, the category ical feature values are converted to their equivalent numerical values using the One Hot Encoding (OHE) technique. shows that all four datasets used in this work have an imbalanced class distribution problem. As such, a Random Over Sampler strategy is used to alleviate this issue. After completing the initial preprocessing, the datasets feature values are scaled using four different FS techniques i.e., QT, PT, Normalizer, and MAS (see their detailed 10 operations in The feature-scaled datasets are then classified using eight different ML classification techniques i.e., AB, RF, DT, KNN, GNB, LR, SVM, and LDA.

DISADVANTAGES

- Imbalanced Class Distribution Handling: Both the existing and proposed systems address the imbalanced class distribution issue through oversampling techniques.
- Mean Value Imputation (MVI): The proposed system replaces MVI with a different imputation method, possibly improving the accuracy of imputed values and reducing potential biases introduced by using mean values.
- One Hot Encoding (OHE): The proposed system replaces OHE with Label Encoding, potentially reducing dimensionality and computational complexity.
- Feature Scaling Techniques: The proposed system simplifies the feature scaling step by using a single technique (possibly Recursive Feature Elimination - RFE) instead of multiple techniques, reducing complexity and potential redundancy.
- Classification Techniques: The proposed system focuses on using a single classification technique (Logistic Regression) instead of multiple techniques, potentially simplifying model selection and evaluation.
- Lack of Model Interpretability: Logistic Regression is generally more interpretable compared to some of the classification techniques used in the existing model,

potentially providing better insights into the factors contributing to ASD prediction.

- Limited Explanation of Preprocessing Steps: The proposed system provides a clearer description of preprocessing steps, mentioning missing values imputation, label encoding, and oversampling, which enhances transparency and reproducibility.
- Potential Overfitting: By simplifying the model pipeline and focusing on a single classification technique, the proposed system may reduce the risk of overfitting, especially if proper regularization techniques are applied during model training.

3.2 PROPOSED SYSTEM

This research aims to create an effective prediction model using different types of ML methods to detect autism in people of different ages. First of all, the datasets are collected, and then the preprocessing is accomplished the missing values imputation, label encoding, and oversampling and create an instance of RFE with the classifier and the desired number of features to select Logistic Regression classification of modeling, performance evaluation, and the results with improved accuracy.

3.3 PROJECT REQUIREMENTS

REQUIREMENTS SPECIFICATION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition associated with brain development that starts early stage of life, impacting a person's social relationships and interaction issues [1], [2]. ASD has restricted and repeated behavioral patterns, and the word spectrum encompasses a wide range of symptoms and intensity [3], [4], [5]. Even though there is no sustainable solution for ASD, simply early difference in a kid's development to focus on improving a child's behaviors and skills in communication [6], [7], [8]. Even so, the identification and diagnosis of ASD are really difficult and sophisticated, using traditional behavioral science. Usually, Autism is most commonly diagnosed at about two years of age and can also be diagnosed later, based on its severity [9], [10], [11]. A variety 11 of treatment strategies are available to detect ASD as quickly as possible. These diagnostic procedures aren't always widely used in practice until a severe chance of developing ASD. The authors in [12] provided a short and observable checklist that can be seen at different stages of a person's life, including toddlers, children, teens, and adults. Subsequently, the authors in [13] constructed the ASD tests mobile apps system for ASD identification as fast as possible, depending on a

range of questionnaire surveys, Q-CHAT, and AQ-10 methods. Consequently, they also created an open-source dataset utilizing mobile phone app information and submitted the datasets to a publicly accessible website called the University of California- Irvine (UCI) machine learning repository and Kaggle for more development in this area of study. Over the past few years, several studies have been conducted incorporating various Machine Learning (ML) approaches to analyze and diagnose ASD and also other diseases, such as diabetes, stroke, and heart failure prediction as quickly as possible [14], [15], [16]. The authors in [17] analyzed the ASD attributes utilizing Rule-based ML (RML) techniques and confirmed that RML helps classification models boost classification accuracy.

3.3.1 HARDWARE REQUIREMENTS

- Hard Disk : 500GB and Above
- RAM : 4GB and Above
- Processor : I3 and Above

3.3.2 SOFTWARE REQUIREMENTS

- Operating System : Windows 10 (64 bit)
- Software : Python 3.7
- Tools : Anaconda (Jupyter Note Book IDE) 3.3

3.3.3 TECHNOLOGIES USED

- Python
- Machine Learning

3.4 DATASET DESCRIPTION

Data Set Name: Autistic Spectrum Disorder Screening Data for Toddlers

Author: Dr Fadi Thabtah

Source: Kaggle

Abstract: Autistic Spectrum Disorder (ASD) is a neurodevelopmental condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across

the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behavior traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. Hence, we propose a new dataset related to autism screening of toddlers that contained influential features to be utilized for further analysis especially in determining autistic traits and improving the classification of ASD cases. In this dataset, we record ten behavioral features (Q-Chat-10) plus other individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behavior science.

Data Type: Predictive and Descriptive: Nominal / categorical, binary and continuous

Task: Classification

Attribute Type: Categorical, continuous and binary

Area: Medical, health and social science

Format Type: Non-Matrix

Does your data set contain missing values? No

Number of Instances (records in your data set): 1054

Number of Attributes (fields within each record): 18 including the class variable

Attributes:

A1-A10: Items within Q-Chat-10 in which questions possible answers : “Always, Usually, Sometimes, Rarly & Never” items’ values are mapped to “1” or “0” in the dataset. For questions 1-9 (A1-A9) in Q-chat-10, if the respose was Sometimes / Rarly / Never “1” is assigned to the question (A1-A9). However, for question 10 (A10), if the respose was Always / Usually / Sometimes then “1” is assigned to that question. If the user obtained More than 3 Add points together for all ten questions. If your child scores more than 3 (Q-chat-10- score) then there is a potential ASD traits otherwise no ASD traits are observed.

Table 1: Details of variables mapping to the Q-Chat-10 screening methods

Variable in Dataset	**Corresponding Q-chat-10-Toddler Features**
A1	Does your child look at you when you call his/her name?
A2	How easy is it for you to get eye contact with your child?
A3	Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)
A4	Does your child point to share interest with you? (e.g. poin9ng at an interes9ng sight)
A5	Does your child pretend? (e.g. care for dolls, talk on a toy phone)
A6	Does your child follow where you're looking?
A7	If you or someone else in the family is visibly upset, does your child show signs of wan9ng to comfort them? (e.g. stroking hair, hugging them)
A8	Would you describe your child's first words as:
A9	Does your child use simple gestures? (e.g. wave goodbye)
A10	Does your child stare at nothing with no apparent purpose?

3.5 INPUT DESIGN

Input design refers to the process of creating the user interface and mechanisms through which users interact with a system or application to input data or commands. It involves designing forms, screens, dialog boxes, and other interface elements that allow users to input information effectively and efficiently.

Input design is a crucial aspect of user interface (UI) and user experience (UX) design, as it directly impacts the usability and functionality of a system. A well-designed input interface should be intuitive, user-friendly, and accessible, minimizing user errors and maximizing productivity.

Fig 3.5.1 Landing Section

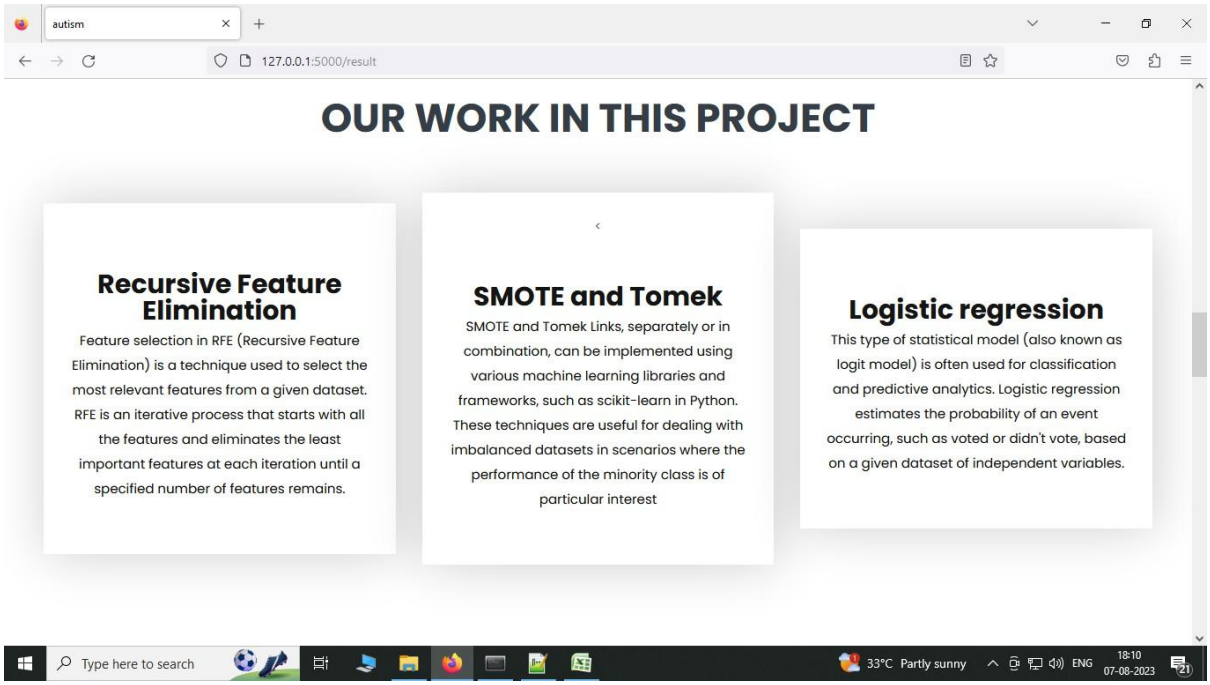


Fig 3.5.2 Details Section

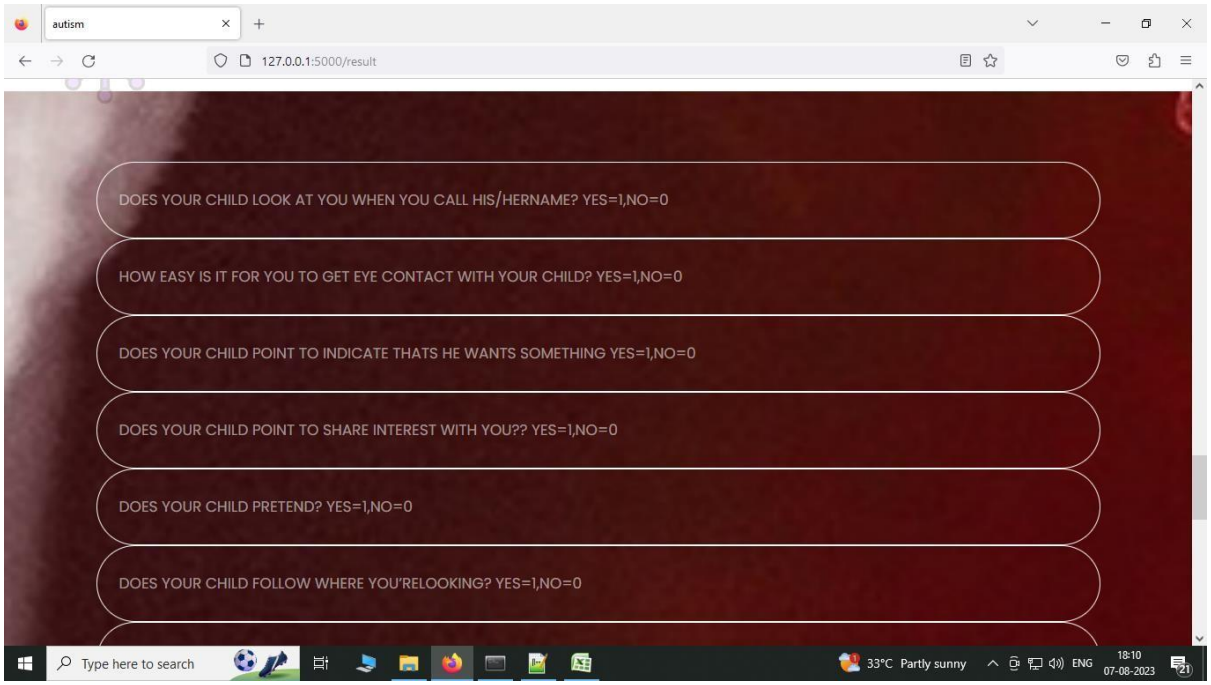


Fig 3.5.3 Data input Section

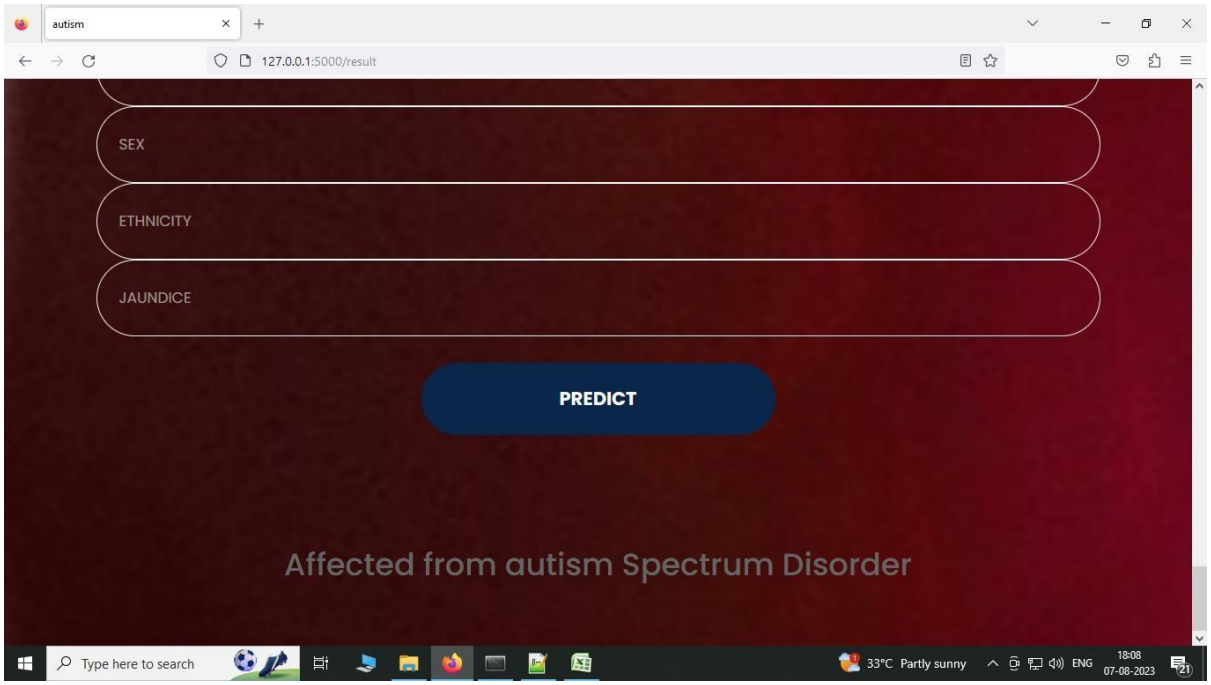


Fig 3.5.4 Result Section

3.6 MODULE DESIGN

3.6.1 Architecture Diagram

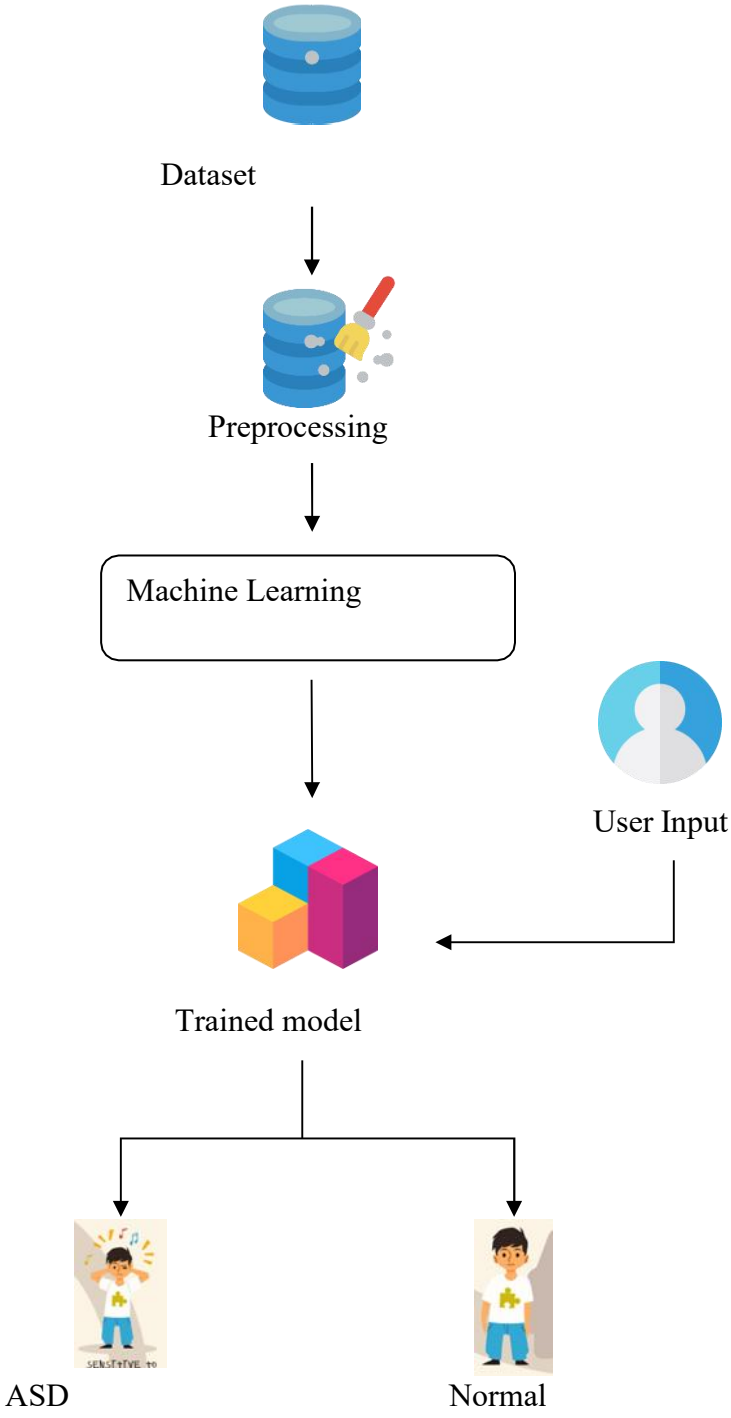


Fig 3.6.1 Architecture Diagram

The early-stage detection of Autism Spectrum Disorders (ASD) presents a significant challenge in clinical settings due to the complexity and variability of symptoms. In this project, we propose a comprehensive machine learning framework aimed at facilitating the early identification of ASD individuals. This framework leverages data preprocessing techniques, machine learning algorithms, and user input to train models capable of classifying individuals as either having ASD or being neurotypical. Herein, we provide a detailed description of the architecture diagram, illustrating the flow of data and processes within the framework.

Architecture Diagram Description:

The architecture diagram represents a structured framework designed for the early detection of Autism Spectrum Disorders (ASD). Each component within the framework plays a crucial role in the process, contributing to the overall efficiency and effectiveness of the detection system.

Dataset:

The process begins with the acquisition of a comprehensive dataset containing relevant features and labels associated with individuals, some of whom have been diagnosed with ASD and others who are neurotypical. This dataset serves as the foundational source of information for subsequent stages.

Preprocessing:

The dataset undergoes preprocessing procedures to ensure that it is clean, standardized, and suitable for analysis. Preprocessing techniques may include data cleaning, normalization, feature engineering, and handling missing values.

Machine Learning:

The preprocessed data is fed into the machine learning component, where various algorithms are employed to learn patterns and relationships between the input features and the corresponding labels. This stage involves model selection, parameter tuning, and evaluation to determine the most suitable approach for ASD detection.

Training Model:

Within this stage, the selected machine learning model is trained using the preprocessed data. The training process involves optimizing the model's parameters to minimize error and enhance predictive performance. Additionally, user input is incorporated to provide flexibility and customization options during the training phase.

User Input:

User input plays a pivotal role in shaping the training process of the model. By allowing users to provide input parameters, preferences, and constraints, the framework ensures adaptability to diverse datasets and clinical requirements. This input is integrated seamlessly into the training pipeline, facilitating collaborative decision-making and refinement of the model.

Trained Model (ASD/Normal):

Upon completion of the training phase, the framework produces trained models capable of classifying individuals into two categories: ASD and normal. These models are equipped with the ability to analyze new data instances and provide predictions regarding the presence or absence of ASD symptoms.

The proposed machine learning framework for early-stage detection of Autism Spectrum Disorders embodies a systematic approach to addressing a critical healthcare challenge. By leveraging advanced data analytics techniques and user-centric design principles, the framework aims to enhance the accuracy, efficiency, and accessibility of ASD diagnosis.

The journey begins with the acquisition of a diverse and representative dataset comprising features relevant to ASD diagnosis, such as behavioral traits, medical history, and demographic information. This dataset serves as the foundation upon which the subsequent stages of preprocessing and analysis are built.

During the preprocessing stage, the raw data undergoes a series of transformations to ensure consistency, reliability, and suitability for machine learning tasks. This may involve standardization of data formats, handling of missing values, and extraction of informative features. By refining the dataset through preprocessing techniques, we aim to optimize the performance of our machine learning models and minimize the risk of bias or errors.

The heart of the framework lies within the machine learning component, where sophisticated algorithms are deployed to extract meaningful patterns and insights from the preprocessed data. Through iterative experimentation and evaluation, we seek to identify the most effective model architecture and hyperparameters for ASD detection. This stage embodies a data-driven approach to decision-making, guided by empirical evidence and statistical analysis.

As we transition to the training model stage, user input emerges as a central element in the model development process. By soliciting feedback, preferences, and domain knowledge from

healthcare professionals and domain experts, we ensure that the trained models are aligned with clinical expectations and requirements. This collaborative approach fosters transparency, accountability, and stakeholder engagement throughout the development lifecycle.

The culmination of our efforts is the generation of trained models capable of discriminating between individuals with ASD and those who are neurotypical. These models serve as valuable tools for healthcare practitioners, aiding in the early identification and intervention of ASD symptoms. By harnessing the power of machine learning and human expertise, we aspire to empower clinicians with actionable insights and decision support tools for improving patient outcomes and quality of life.

UML DIAGRAMS

Unified Modeling Language (UML) is a general purpose modelling language. The main aim of UML is to define a standard way to visualize the way a system has been designed. It is quite similar to blueprints used in other fields of engineering.

3.6.2 USECASE DIAGRAM

Use case diagrams are considered for high level requirement analysis of a system. When the requirements of a system are analyzed the functionalities are captured in use cases. So, it can say that uses cases are nothing but the system functionalities written in an organized manner.

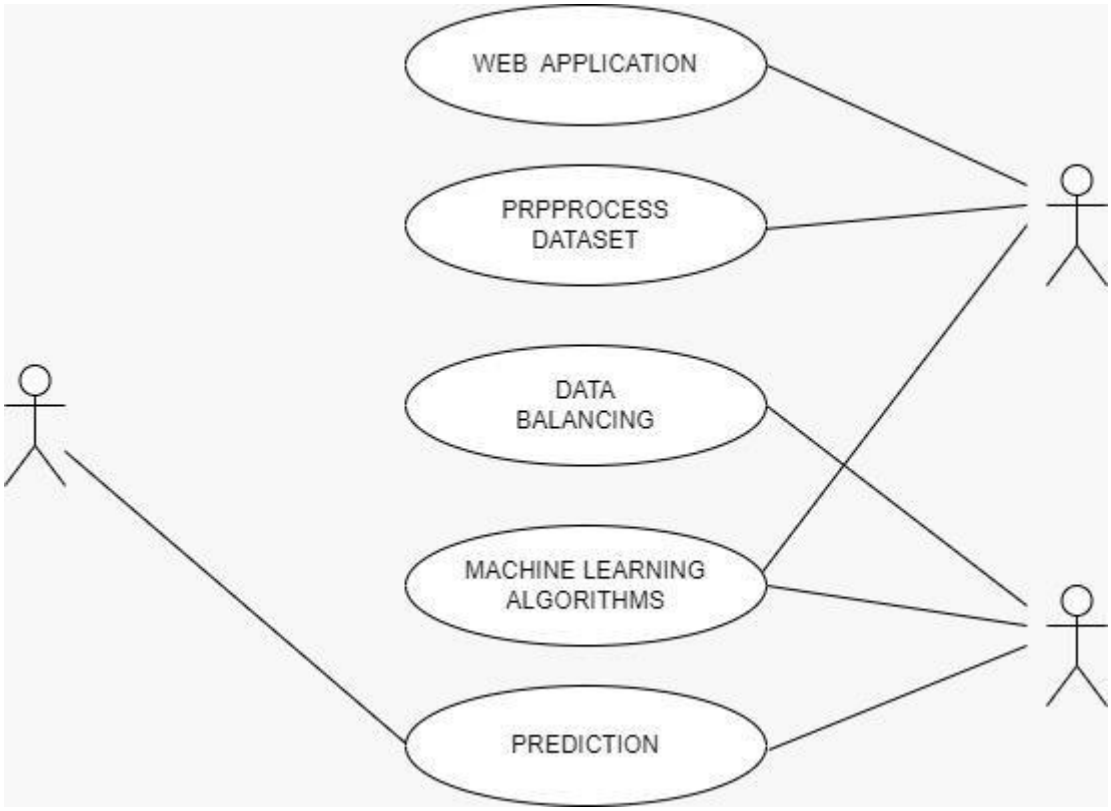
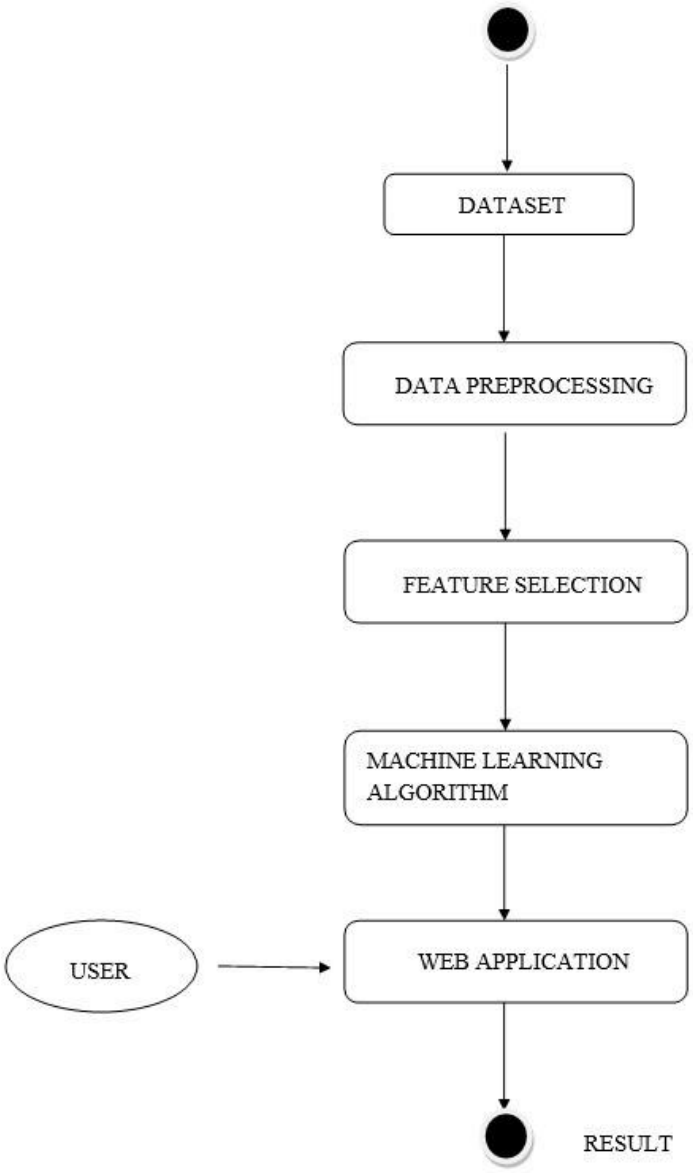


Fig 3.6.2 Usecase Diagram

3.6.3 ACTIVITY DIAGRAM

A graphical representation of an executed set of procedural system activities and considered a state chart diagram variation. Activity diagrams describe parallel and conditional activities, use cases and system functions at a detailed level.

Fig 3.6.3 Activity Diagram



3.6.4 COLLABORATION DIAGRAM

UML Collaboration Diagrams illustrate the relationship and interaction between software objects. They require use cases, system operation contracts and domain model to already exist. The collaboration diagram illustrates messages being sent between classes and objects.

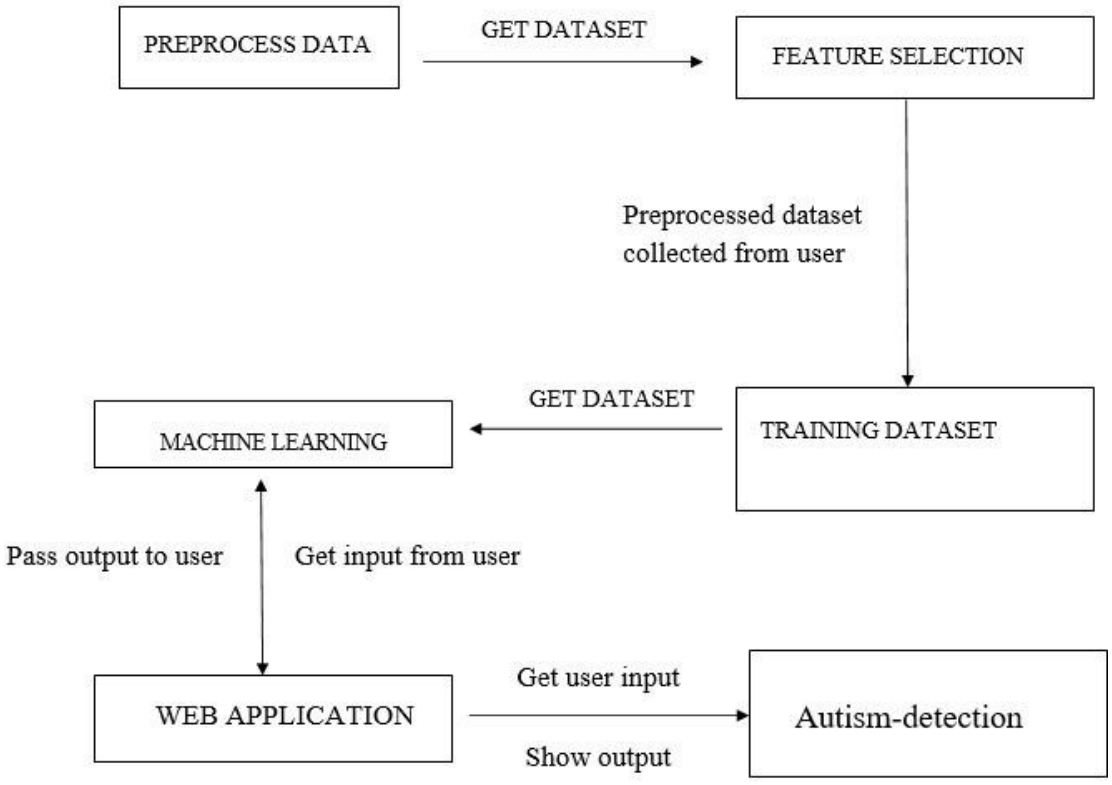


Fig 3.6.4 Collaboration Diagram

3.6.5 DATA FLOW DIAGRAM

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. It can be used for the visualization of data processing (structured design).Data flow diagrams are also known as bubble charts. DFD is a designing tool used in the top down approach to Systems Design. DFD levels are numbered 0, 1 or 2, and occasionally go to even Level 3 or beyond. DFD Level 0 is also called a Context Diagram.

Level 0:

Level 1:

Level 2:

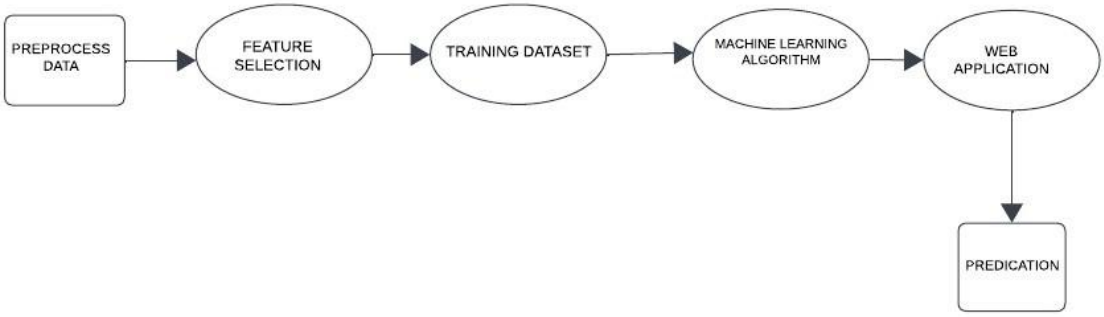


Fig 3.6.5 DFD Diagram

CHAPTER-4
SYSTEM IMPLEMENTATION

CHAPTER 4

4.1 MODULES

- Data Pre-Processing
- Algorithm Implementation
- Prediction

4.2 MODULE EXPLANATION:

Data Pre-Processing:

We collect the four ASD datasets (Toddlers, Adolescents, Children, and Adults) from the publicly available repositories: Kaggle and UCI ML the ASD Tests smartphone app for Toddlers, Children, Adolescents, and Adults ASD screening using QCHAT-10 and AQ-10. The application computes a score of 0 to 10 for every individual, with which the final score is 6 out of 10 which indicates an individual has positive ASD. In addition, ASD data is obtained from the ASD Tests app while open-source databases are developed in order to facilitate research in this area. The detailed description of the Toddlers, Children, Adolescents, and Adults ASD datasets

Algorithm Implementation:

The Classification Algorithms to produce the best results. We are using KNN, Logistic regression and Random Forest Algorithm to predict the ASD disease using ML. On an analysis conducted within various algorithms, the Logistic Regression was found to provide highest efficiency. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate of error.

- **Naive Bayes Classifier:** In Naive bayes exper- iment we used the Gaussian Naive bayes which showed some promising results and the results were better than Supporting Vector Machine classifier. We kept the parameters for Naive Bayes as default. We did not apply any different parameters for Naive bayes as it showed some good results in default parameters. It gave 89% of accuracy which is better thanthe first algorithm and the precision score is 100% which is impressive. Also, Recall score was 84% which is not so good as before and the F1 score it showed is 91%. It took 1.53 seconds to complete the experiment and to produce the result which is the lowest

among all other algorithms.The core formula used in Naive Bayes classification to predict the class (C) with the highest probability for a new data point (X) is:

$$P(C \mid X) = (P(X \mid C) * P(C)) / P(X)$$

where:

- P(C | X) represents the probability of class C occurring given the data point X (what we want to predict)
- P(X | C) represents the probability of observing data point X given that it belongs to class C
- P(C) represents the prior probability of class C (how frequent it is in the training data)
- P(X) represents the total probability of observing data point X (usually a constant value for normalization)

- **Logistic Regression:** Logistic regression is a machine learning method for categorizing things. It analyzes data to predict the probability of an event happening (like spam filtering or disease detection) by transforming the relationships between features and outcomes into a probability between 0 and 1. While it requires a linear relationship between features and the outcome, it's easy to interpret and implement, making it a popular choice for various classification tasks.

$$p(y = 1) = 1 / (1 + e^{(-z)})$$

where:

- **p(y = 1)** represents the predicted probability of the positive outcome (often denoted by the Greek letter sigma, σ).
- **e** is the base of the natural logarithm (approximately 2.718)
- **z** is the weighted sum of the independent variables (features) and their corresponding coefficients, often referred to as the **linear predictor**. It can be expressed as:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where:

- β_0 is the bias term (coefficient for the constant term)
- β_i are the coefficients for each independent variable (x_i)

This formula essentially calculates the log-odds of the event happening and then uses the logistic function (1 / (1 + e^(-z))) to convert it into a probability between 0 and 1. The coefficients (β) determine the influence of each feature on the probability.

Support Vector Machine: Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates the data into different classes while maximizing the margin between the classes.

Prediction:

Building a robust machine learning model requires a well-prepared dataset. One crucial step in this preparation is splitting the data into training and testing sets. This passage will explore how to achieve this split using your dataset of 1456 data points.

We'll utilize a train-test split approach, where a portion of the data is held out for testing, often referred to as the holdout or test set. In this case, we'll allocate 15% (or a split ratio of 0.15) of the data to the test set, resulting in a size of approximately 219 data points (calculated by multiplying the total sample size of 1456 by the split ratio). The remaining 85% (1237 data points) will constitute the training set.

This split ensures the model doesn't become overly familiar with the training data, a phenomenon known as overfitting. The test set allows for a more unbiased assessment of the model's performance on unseen data. By effectively splitting your dataset, you'll lay the groundwork for a more reliable and generalizable machine learning model.

CHAPTER-5
PERFORMANCE ANALYSIS

CHAPTER 5
PERFORMANCE ANALYSIS

Precision:

Precision tells us how many of the items our model predicted as positive are truly positive. A higher precision value indicates that the model has fewer false positives and is more accurate in identifying positive instances.

Precision is calculated using the following formula:

Precision=True Positives / (True Positives + False Positives)

Where:

- True Positives (TP) are the instances that were predicted as positive by the model and are actually positive according to the ground truth.
- False Positives (FP) are the instances that were predicted as positive by the model but are actually negative according to the ground truth.

Recall:

Recall (also known as sensitivity or true positive rate) is a measure of the ability of the model to correctly identify all positive instances in the dataset.

Recall is calculated using the following formula:

Recall= True Positives / (True Positives + False Negatives)

Where:

- True Positives (TP) are the instances that were predicted as positive by the model and are actually positive according to the ground truth.
- False Negatives (FN) are the instances that were predicted as negative by the model but are actually positive according to the ground truth.

Accuracy:

Accuracy measures the proportion of correctly classified instances among all instances in the dataset. It provides an overall assessment of the model's performance.

Accuracy is calculated using the following formula:

Accuracy=Number of Correct Predictions / Total Number of Predictions

F1 Score:

The F1 score is a measure of a classification model's accuracy that considers both the model's precision and recall. It is the harmonic mean of precision and recall, providing a single metric that balances both aspects of the model's performance.

The F1 score is calculated using the following formula:

F1=2 × (Precision × Recall / Precision + Recall)

Where:

- Precision is the ratio of true positives to the sum of true positives and false positives.
- Recall is the ratio of true positives to the sum of true positives and false negatives.

Confusion Matrix:

A confusion matrix is a table that is often used to evaluate the performance of a classification model. It summarizes the performance of a classification algorithm by comparing the actual values of the target variable with the values predicted by the model.

A confusion matrix is typically organized into four quadrants:

1. True Positives (TP): These are the instances where the model correctly predicts the positive class.
2. True Negatives (TN): These are the instances where the model correctly predicts the negative class.
3. False Positives (FP): These are the instances where the model incorrectly predicts the positive class when it is actually negative (Type I error).
4. False Negatives (FN): These are the instances where the model incorrectly predicts the negative class when it is actually positive (Type II error).

Table 2: Performance analysis table

Algorithm	Precision	Recall	Accuracy	F1 Score
Logistic Regression	0.99	0.99	0.99	0.99
Naïve bayes	0.945	0.981	0.915	0.946
Support Vector Machine	0.181	0.054	0.398	0.083

Confusion Matrix of Logistic Regression:

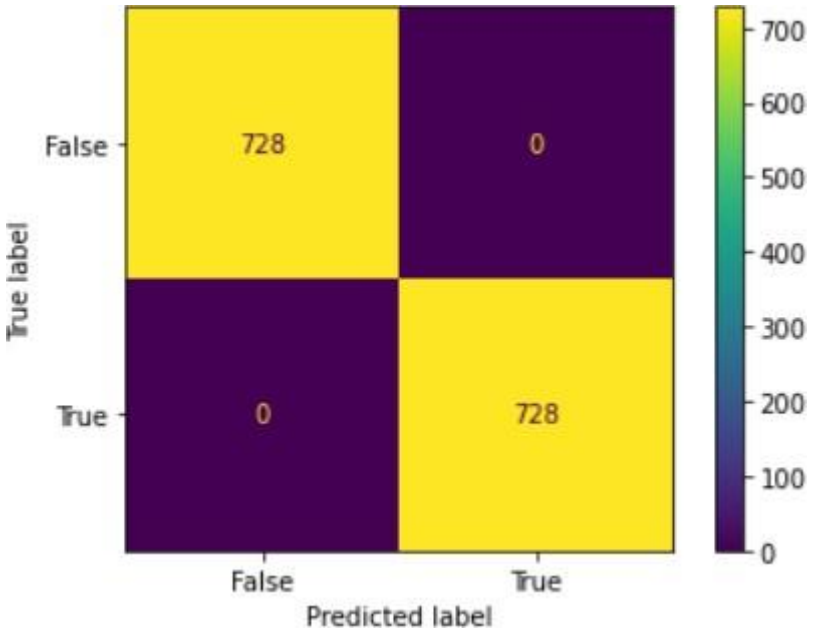


Fig 5.1 Confusion Matrix of Logistic Regression

Confusion Matrix of Naïve Bayes:

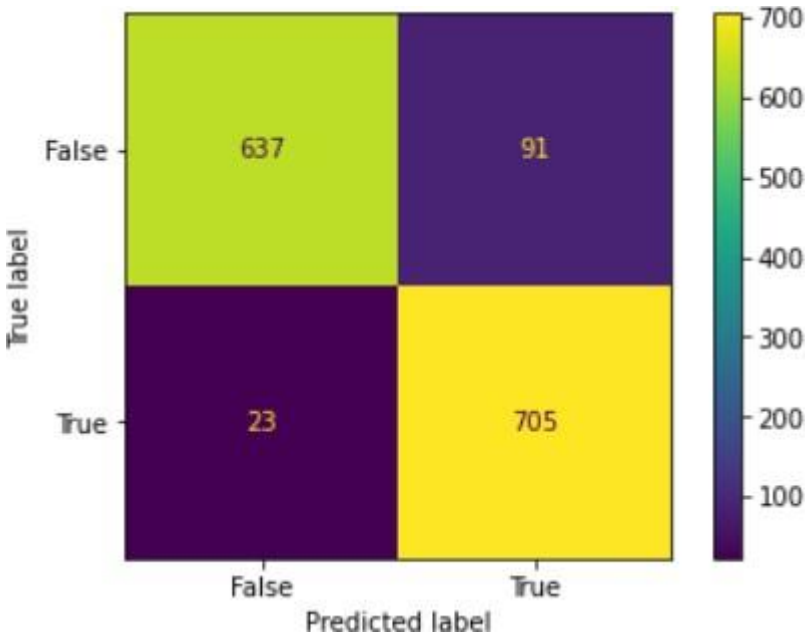


Fig 5.2 Confusion Matrix of Naïve Bayes

Confusion Matrix of Support Vector Machine:

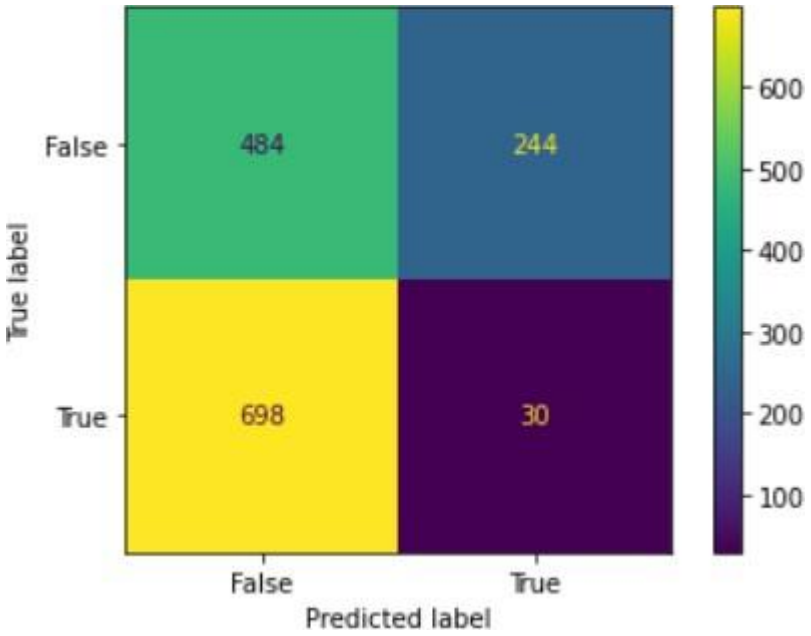


Fig 5.3 Confusion Matrix of SVM

CHAPTER-6

CONCLUSION AND FUTURE WORK

CHAPTER 6

CONCLUSION AND FUTURE WORK

Our proposed model focuses on enhancing early detection of autism while preserving privacy through thoughtfully designed parent-questionnaires. Utilizing datasets from Q-CHAT and AQ tools, we employed supervised learning algorithms including Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. Our model achieved notable accuracy rates for toddlers: SVM (39%), Logistic Regression (99%), and Naive Bayes (91%).

However, a significant limitation arose from the insufficiently large dataset for training. This constraint resulted in overfitting during the implementation of certain algorithms, where the model fit the training data too closely but struggled to generalize to new data. Consequently, we had to discontinue the use of some algorithms due to this modeling error.

While prior research has addressed autism detection across various age ranges, our focus on early ASD detection, particularly in toddlers, offers a novel approach crucial for improving outcomes. By concentrating on toddlers, we aim to provide more precise results and facilitate earlier interventions.

To overcome these limitations, future efforts will concentrate on gathering more diverse datasets from multiple sources to enhance the model's accuracy. Additionally, we plan to develop a user-friendly mobile application based on our model, enabling individuals to effortlessly assess early autism symptoms and seek professional help when necessary. This approach aims to the delays and financial burdens associated with traditional diagnosis methods.

In summary, our proposed model offers the potential to guide individuals at a very early age, preventing situations from worsening and reducing the costs associated with delayed diagnosis.

APPENDICES

APPENDICES

A.1 SDG GOALS

Goal 3

Ensuring Good Health and Well-being: Focuses on reducing premature deaths caused by non-communicable diseases, which includes mental health conditions such as ASD. This involves raising awareness, improving detection, and enhancing treatment for such conditions.

Goal 4

Providing Quality Education: Aims to guarantee equal access to education and vocational training for vulnerable groups, including people with disabilities like ASD. Detecting ASD early enables timely intervention and support, ultimately improving educational outcomes for affected individuals.

Goal 10

Promoting Reduced Inequalities: Highlights the importance of empowering and socially including all individuals, regardless of their background or status. This entails ensuring that people with ASD have equitable access to healthcare, education, and employment opportunities.

A.2 SOURCE CODE

Pre processing:

```
import pandas as pd
import numpy as np
data=pd.read_csv('./dataset.csv')
data.shape
data.describe()
data.columns
data=data.rename(columns={'Class/ASD Traits ':'class','Who completed the
test':'Who_completed_the_test','Qchat-10-Score':'Qchat_10_Score'}
data.isna().sum()
data['class'].value_counts()
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10,10))
```

```
sns.countplot(x='class',data=data)
numarical_columns = [col for col in data.columns if data[col].dtype == 'int64']
numarical_columns = data[numarical_columns].copy()
numarical_columns
categorical_columns = [col for col in data.columns if data[col].dtype == 'object']
categorical_columns = data[categorical_columns].copy()
categorical_columns
categorical_columns['Ethnicity'].unique()
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()


for column in categorical_columns.columns:
    categorical_columns[column] =
label_encoder.fit_transform(categorical_columns[column])
categorical_columns
inner_join_result = pd.merge(numarical_columns, categorical_columns,
on=data['Case_No'], how='inner')
```

```
data=inner_join_result
data
X=data.drop(columns='class')
Y=data['class'] print(X.shape)
print(Y.shape)
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression

# Assuming you have already loaded your cervical cancer dataset into X and y

# Create an instance of the classifier you want to use
classifier = LogisticRegression()

# Create an instance of RFE with the classifier and the desired number of features to
select
rfe = RFE(estimator=classifier, n_features_to_select=14) # Adjust the number of features
as needed

# Perform RFE
selected_features = rfe.fit_transform(X, Y)

# Retrieve the selected feature indices
feature_indices = rfe.get_support(indices=True)

# Print the selected feature indices and their names
print("Selected Features:")
for index in feature_indices:
    print(f"- {X.columns[index]}")
selected_columns =
['A1','A2','A3','A4','A5','A6','A7','A8','A9','A10','Qchat_10_Score','Sex','Ethnicity','Jaundic
e','class']
data = data[selected_columns]
data.to_csv('./preprocess.csv',index=False)
```

Model training:

```
import pandas as pd
import numpy as np
data=pd.read_csv('./preprocess.csv')
data['class'].value_counts()
import matplotlib.pyplot as plt
import seaborn as sns
sns.countplot(x='class',data=data)
X=data.drop(columns='class')
Y=data['class']
from imblearn.combine import SMOTETomek
sm = SMOTETomek()
X_bal, Y_bal = sm.fit_resample(X, Y)
import seaborn as sns
sns.countplot(x=Y_bal,data=data)
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X_bal, Y_bal, test_size=0.15,
random_state=111)
print (X_train.shape)
print (X_test.shape)
print (Y_train.shape)
print (Y_test.shape)
from sklearn.metrics import confusion_matrix, accuracy_score, ConfusionMatrixDisplay
from sklearn.linear_model import LogisticRegression
lrc = LogisticRegression(solver='liblinear', penalty='l1')
lrc.fit(X_train,Y_train)
LR=lrc.score(X_train,Y_train)
test_accuracy = lrc.score(X_test, Y_test)
print('Score:{}'.format(test_accuracy))
Y_pred = lrc.predict(X_test)
confusion_mat = confusion_matrix(Y_test,Y_pred)
print("Confusion Matrix")
print(confusion_mat)
trsc = accuracy_score(Y_test,Y_pred)
```

```
cm_display = ConfusionMatrixDisplay(confusion_matrix = confusion_mat,
display_labels = [False, True])
cm_display.plot()
plt.show()
from sklearn.naive_bayes import MultinomialNB
mnb = MultinomialNB(alpha=0.2)
mnb.fit(X_train,Y_train)
MNB=mnb.score(X_train,Y_train)
test_accuracy = mnb.score(X_test, Y_test)
print('Score:{}'.format(test_accuracy))
Y_pred = mnb.predict(X_test)
confusion_mat = confusion_matrix(Y_test,Y_pred)
print("Confusion Matrix")
print(confusion_mat)
trsc = accuracy_score(Y_test,Y_pred)
cm_display = ConfusionMatrixDisplay(confusion_matrix = confusion_mat,
display_labels = [False, True])
cm_display.plot()
plt.show()
from sklearn.svm import SVC
svc = SVC(kernel='sigmoid', gamma=1.0)
svc.fit(X_train,Y_train)
SVC=svc.score(X_train,Y_train)
test_accuracy = svc.score(X_test, Y_test)
print('Score:{}'.format(test_accuracy))
Y_pred = svc.predict(X_test)
confusion_mat = confusion_matrix(Y_test,Y_pred)
print("Confusion Matrix")
print(confusion_mat)
trsc = accuracy_score(Y_test,Y_pred)
cm_display = ConfusionMatrixDisplay(confusion_matrix = confusion_mat,
display_labels = [False, True])
cm_display.plot()
plt.show()
```

```
lr=LR*100
mnb=MNB*100
svc=SVC*100
height=[lr,mnb,svc]
bars=['logistic','MultinomialNB','svc']
x_pos=np.arange(len(bars))
plt.bar(x_pos, height, color=['#69C96E', '#4482C1', '#B24BF3'])
plt.show
import joblib
joblib.dump(lrc,'./model.jlb')
```

A.3 SCREENSHOT

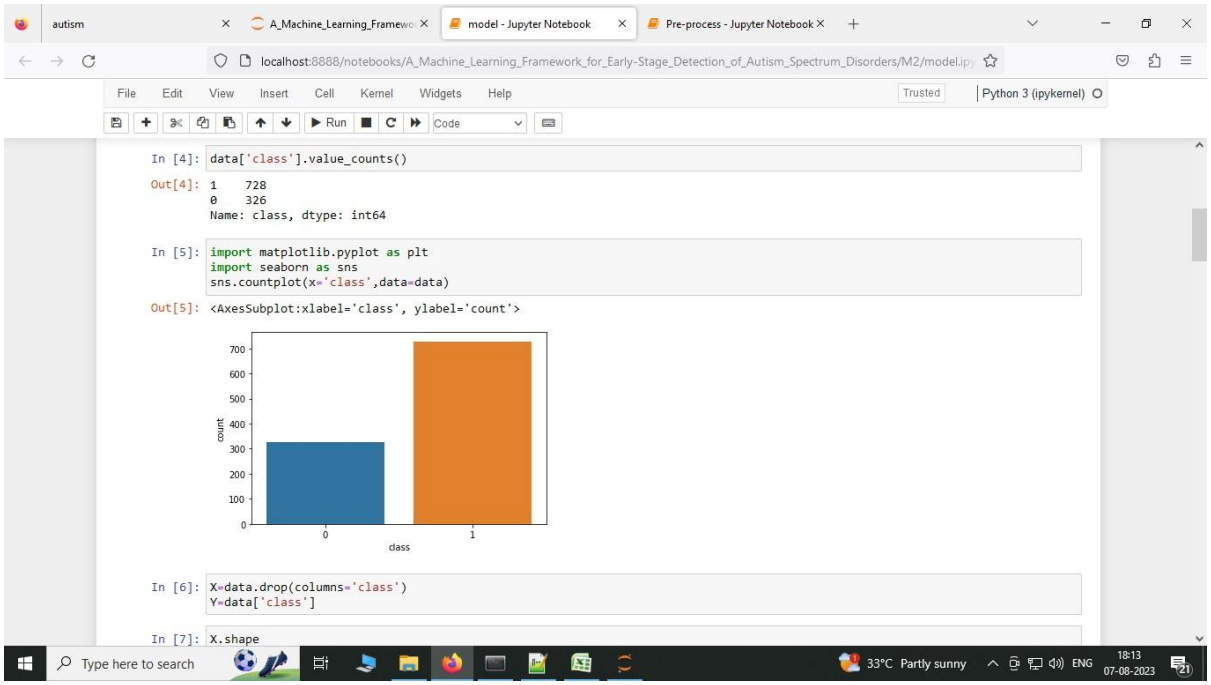


Fig A.3.1 Screenshot of data plot 1

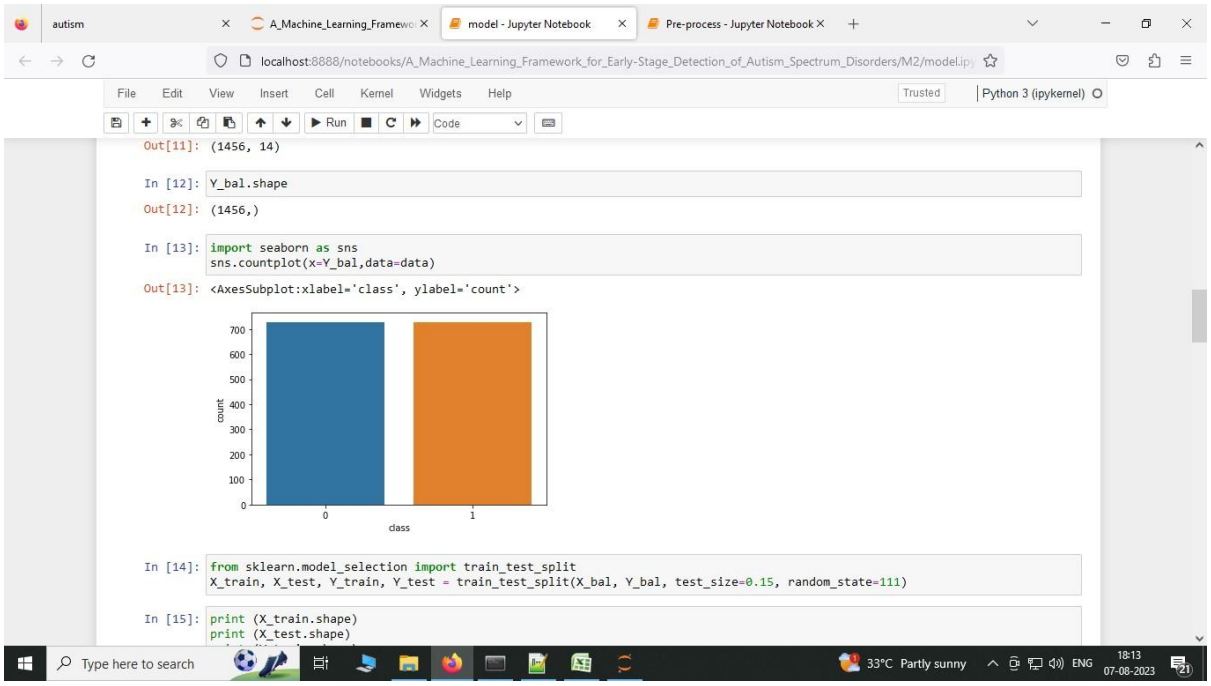


Fig A.3.2 Screenshot of data plot 2

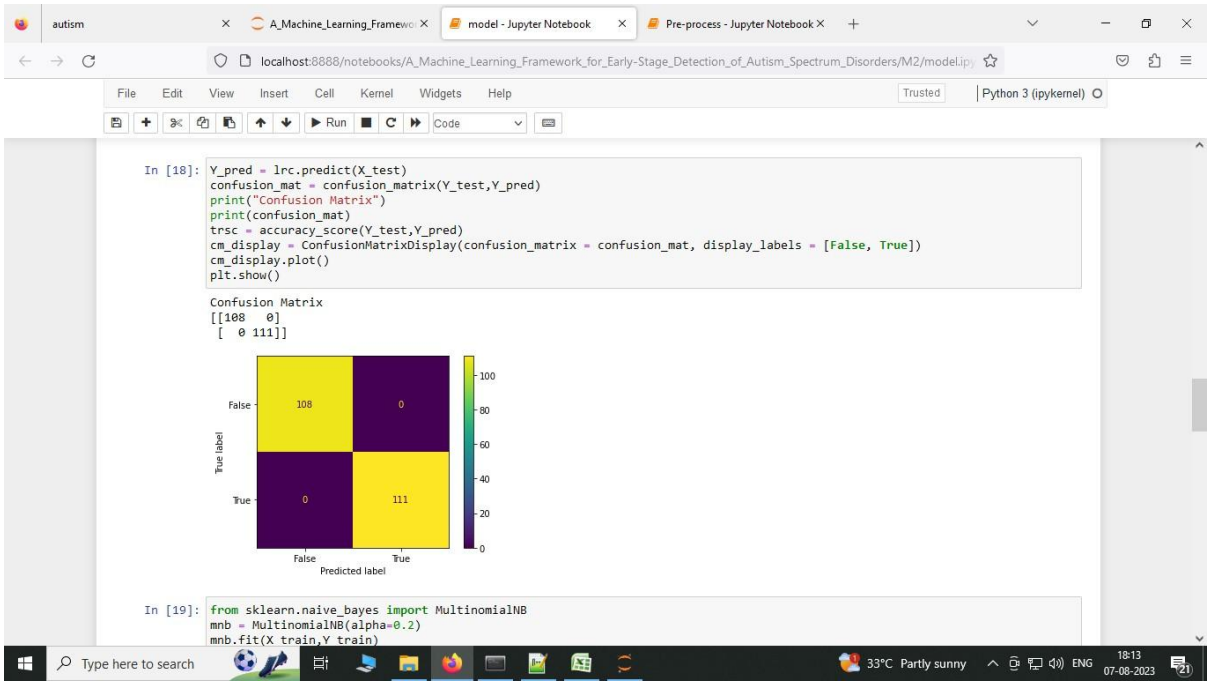


Fig A.3.3 Screenshot of Confusion matrix (LR)

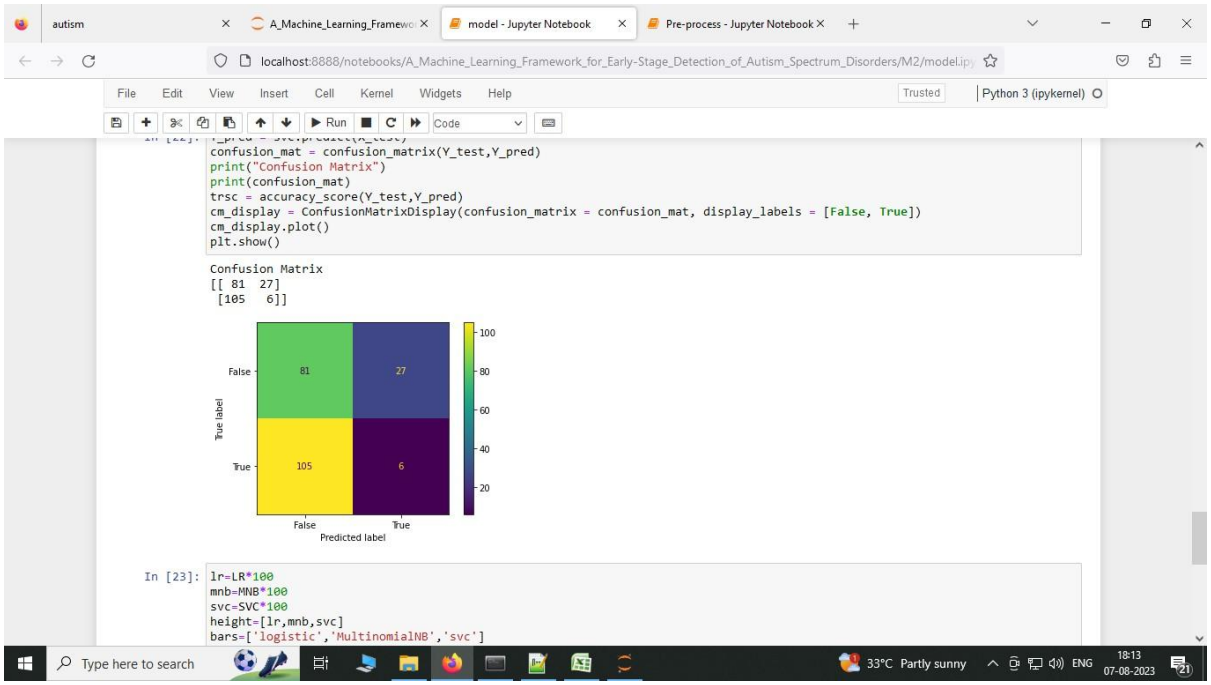


Fig A.3.4 Screenshot of Confusion matrix (SVM)

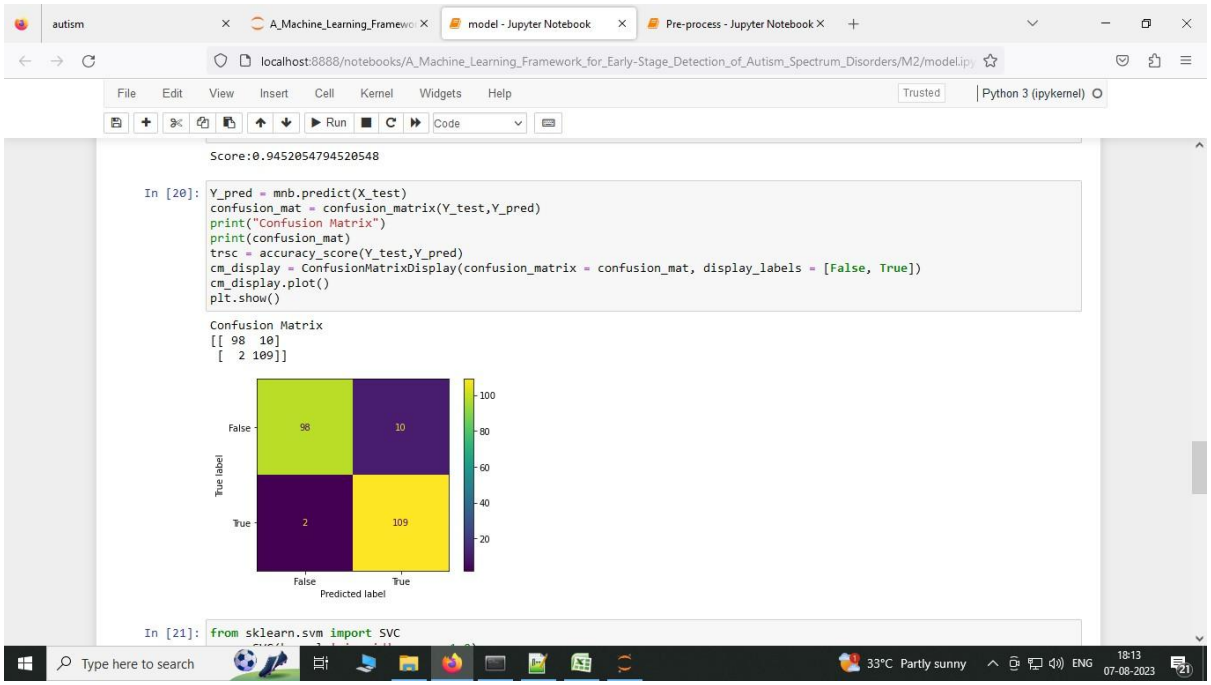


Fig A.3.5 Screenshot of Confusion matrix (NB)

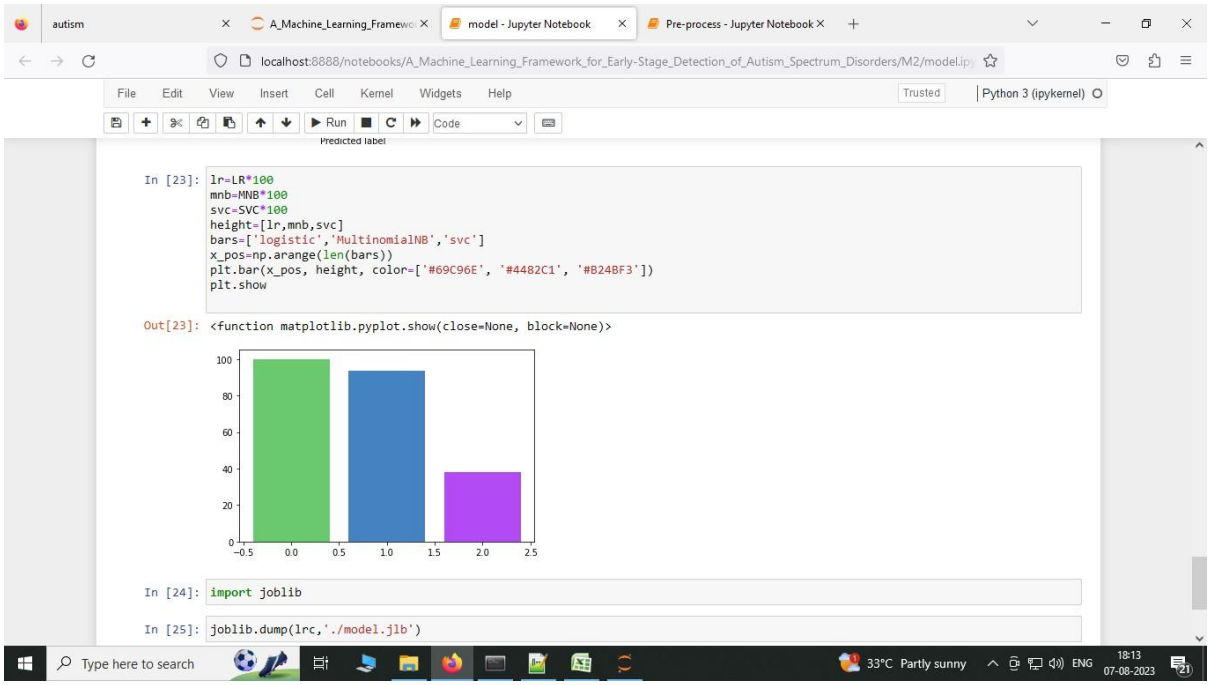


Fig A.3.6 Screenshot of Accuracy plot

Fig A.3.7 Screenshot of landing section

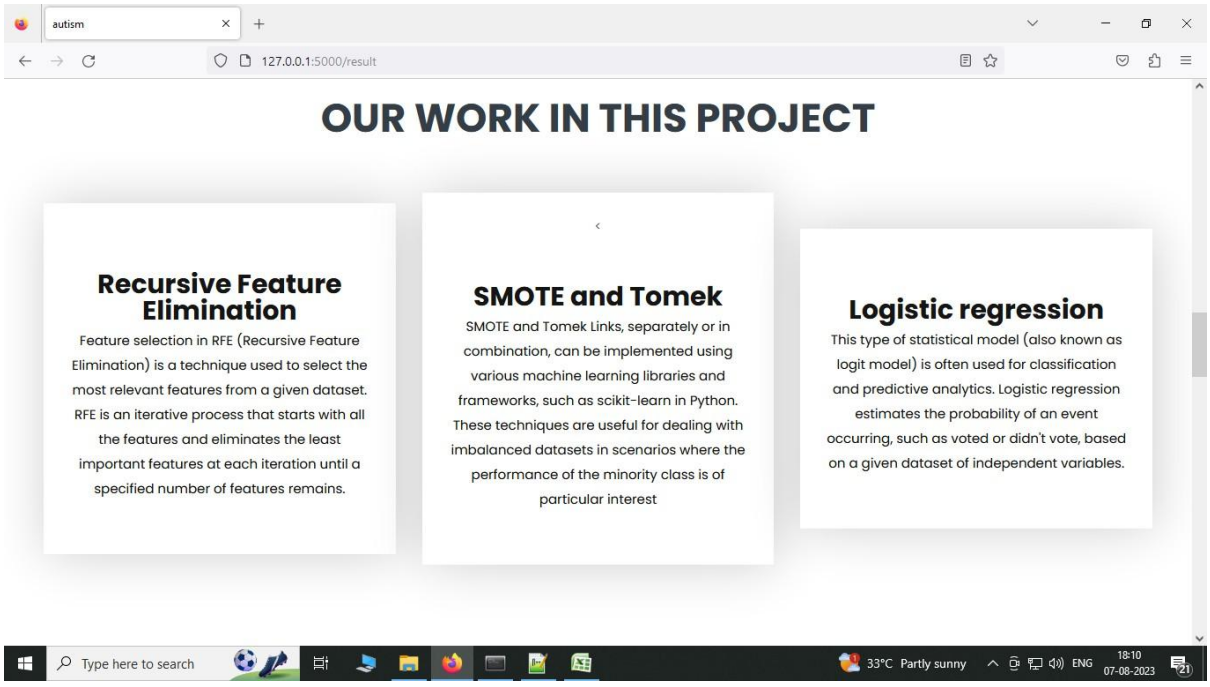


Fig A.3.8 Screenshot of details section

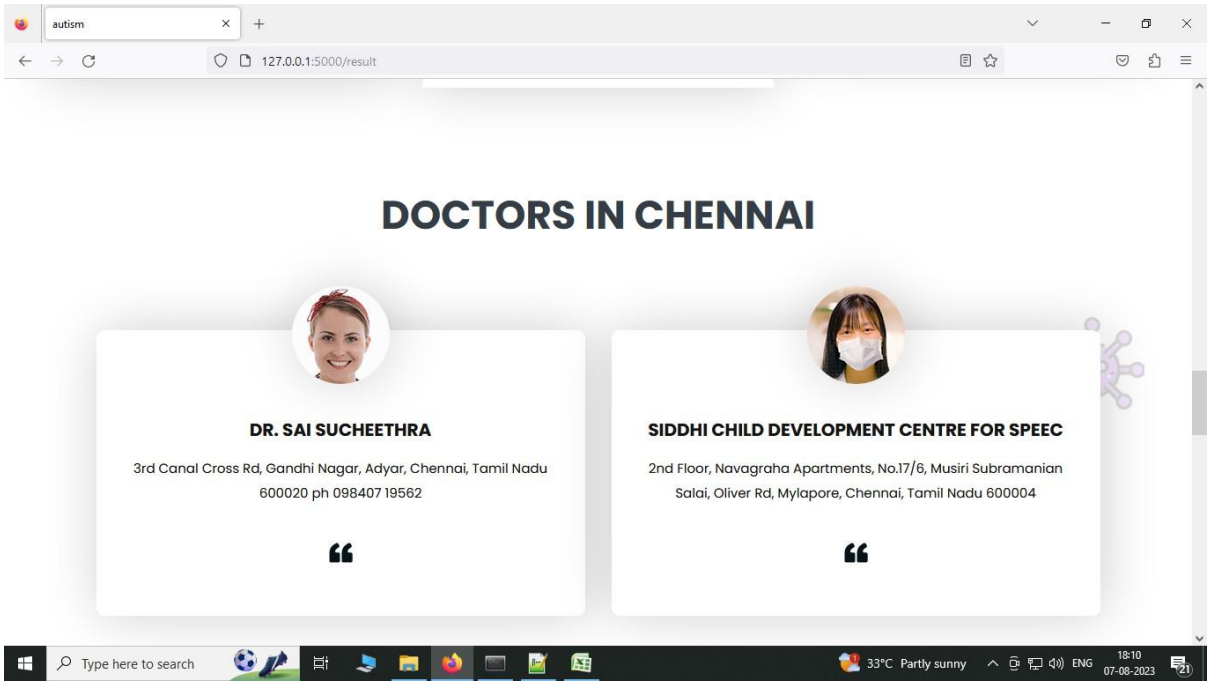


Fig A.3.9 Screenshot of doctors section

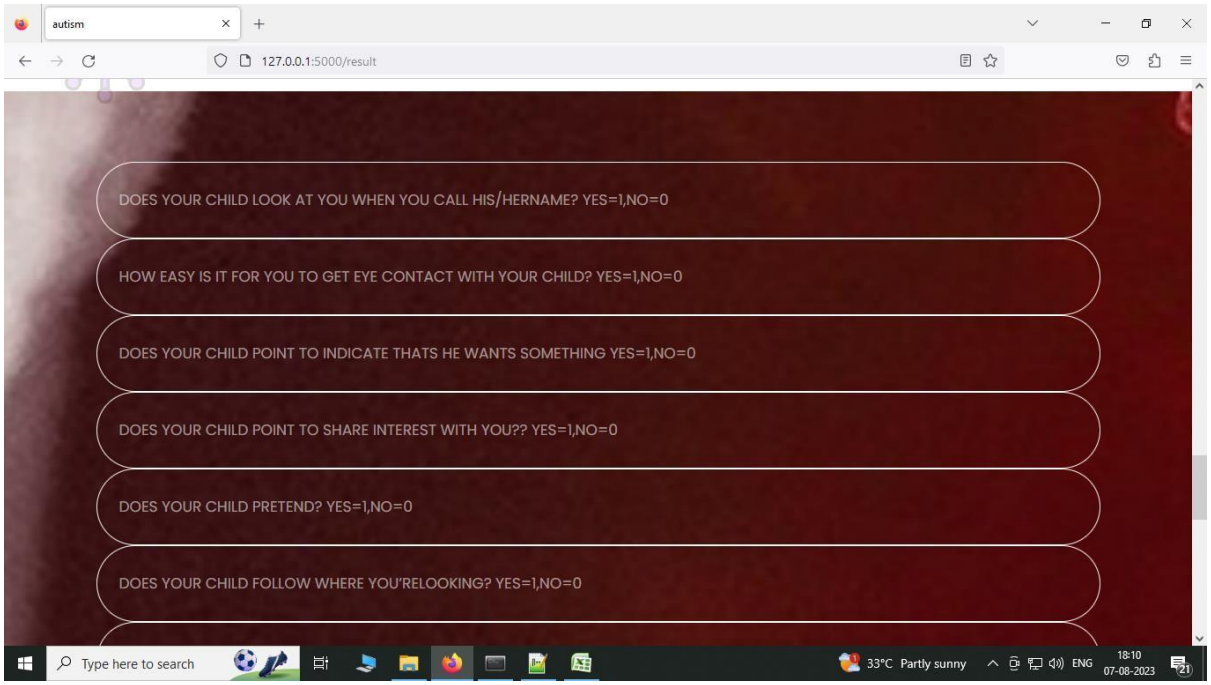


Fig A.3.10 Screenshot of data input section

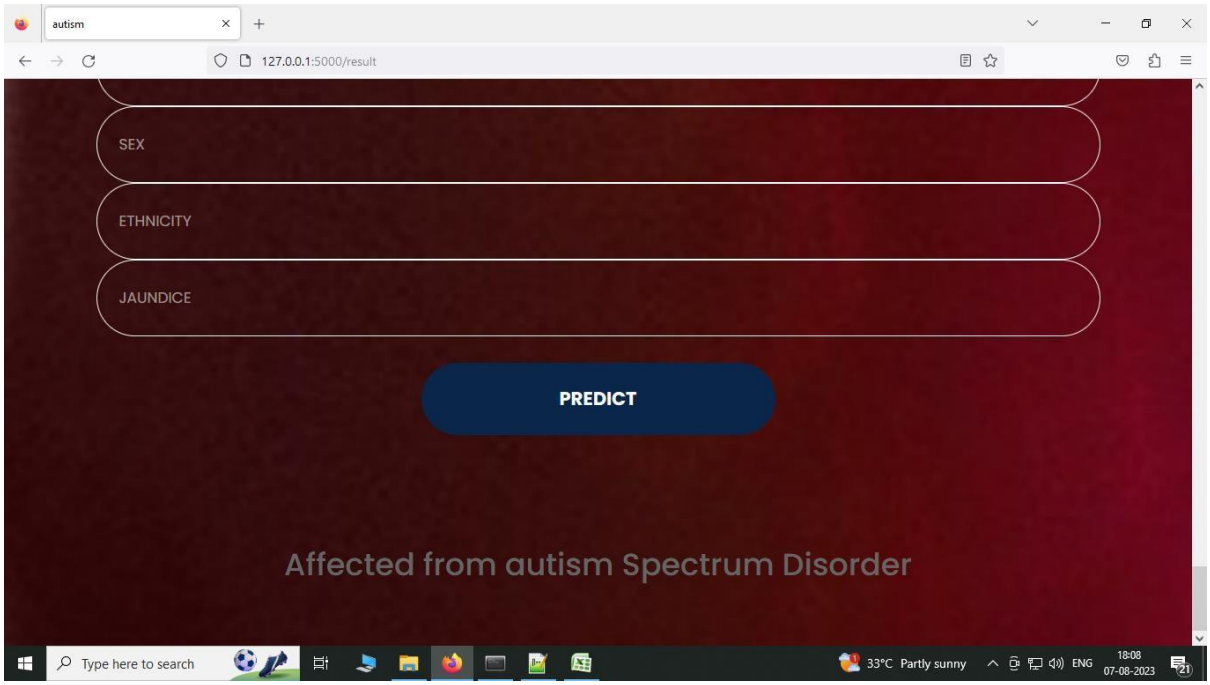


Fig A.3.11 Screenshot of data output section

A.4 Plagiarism Report

A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders.

Krishnaraghavan M
Department of Computer Science And Engineering
Panimalar Engineering College
Chennai, India
krishnaraghavanm@gmail.com

Kaushik S
Department of Computer Science And Engineering
Panimalar Engineering College
Chennai, India
kaushiksathyanath383@gmail.com

Manigandan A
Department of Computer Science And Engineering
Panimalar Engineering College
Chennai, India
manigandan20010103@gmail.com

Dr.L.Jabasheela,M.E.,Ph.D.,
Head of the Department CSE
Panimalar Enginering College
Chennai, India
csehod@panimalar.ac.in

***Abstract* - Individuals who suffer from autism spectrum disorder (ASD) have numerous challenges since it impedes their ability to engage and communicate with others. With one in every 59 children classified as having ASD, the incidence of the disorder is alarming. Effective therapy requires early detection, but many children do not receive a diagnosis until later in life, which makes receiving healthcare more difficult. Identifying this gap, our research uses machine learning approaches to increase the precision and speed of ASD diagnoses. Through gathering large amounts of surveillance footage and crafting focused questions, we hope to improve the accuracy of diagnosis. By utilizing supervised learning algorithms, specifically Logistic Regression, Naive Bayes and Support Vector Machine (SVM), we are able to identify between children with autism and those without it with remarkable precision and effectiveness. Our ultimate goal is to create an internet tool that can accurately and early identify ASD, lowering healthcare expenses and improving the lives of those who are impacted.**

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental contamination characterised through continual demanding situations with social interaction, communication, and conduct repetition [1]. Over the past few decades, the frequency of ASD has increased significantly; according to contemporary estimates, one in 59 children globally has an ASD diagnosis [2]. It is generally acknowledged that a mix of hereditary and environmental variables plays a role in the development of ASD, even though its precise causes are still unknown [3].

Even with great progress in our knowledge of ASD, difficulties with early identification and treatment still exist. Many people with ASD do not receive a diagnosis until much later in childhood or even as adults, which delays their access to necessary support services and interventions [4]. Early diagnosis of ASD is critical because it allows for the timely implementation of evidence-based therapies, which have the potential to greatly enhance the quality of life and outcomes for those with ASD [5]. In recent years, there has been an increase in the use of machine learning techniques to assist in the early diagnosis and detection of ASD.

These methods show promise for evaluating big datasets, spotting subtle trends, and producing precise forecasts based on various behavioral and clinical characteristics [6]. Using machine learning, researchers hope to create strong frameworks that will help identify ASD in its early stages, allowing for earlier therapies and better long-term results.

A thorough machine learning framework for the early diagnosis of autism spectrum disorders is proposed in this research. By utilizing a wide range of data sources and sophisticated algorithms, our framework seeks to increase the precision and efficacy of ASD diagnosis, which will ultimately lead to better clinical practice and results for ASD sufferers.

II. EXISTING MODEL

The goal of the current research is to develop a reliable prediction model for the early identification of autism spectrum disorders in different age groups by employing a variety of machine learning (ML) techniques. First, extensive datasets are compiled that include pertinent characteristics suggestive of autism spectrum disorders (ASD) across various populations. The datasets are then carefully refined using preprocessing procedures to provide the best possible quality for the analysis that follows. Preprocessing includes a number of crucial procedures, such as feature encoding, imputation of missing values, and the use of oversampling techniques to rectify class imbalance.

The Mean Value Imputation (MVI) technique is utilized to handle missing values in the dataset. This allows for the estimation of missing data points by utilizing the mean value of observed instances within the dataset. Furthermore, the One Hot Encoding (OHE) technique converts the values of categorical features into numerical equivalents, making it easier to incorporate categorical variables into the machine learning model. Moreover, a Random Over Sampler approach is employed to address the issue of imbalanced class distributions in the dataset, guaranteeing sufficient representation of minority classes during model training.

After preprocessing, the datasets are feature-scaled using missing values imputation, label encoding, and oversampling and create an instance of RFE with the classifier and the desired number of features to select Logistic Regression classification of modeling, performance evaluation, and the results with improved accuracy. By standardizing the range of feature values, these strategies hope to improve the performance of later classification systems. The feature-scaled datasets are then classified using a variety of machine learning (ML) classification methods, such as Naive Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM).

The most efficient methodology for early-stage ASD detection can be identified with the use of this holistic approach, which allows for thorough examination and comparison of the performance of several ML algorithms in ASD detection.

III. PROPOSED SYSTEM

The main goal of this study is to create a reliable prediction model for the early identification of autism spectrum disorders (ASD) in a range of age groups by applying several machine learning (ML) techniques. This objective is met by taking a multi-step method. First, large-scale datasets are collected that include a variety of clinical and demographic characteristics related to the diagnosis of ASD. Preprocessing is then carried out on the gathered data to make sure it is appropriate for analysis.

A number of crucial actions are taken during the preparation stage with the goal of improving the dataset's consistency and quality. Imputation techniques are employed to handle missing values in the dataset, guaranteeing that no significant information is lost as a result of incomplete data. Moreover, label encoding techniques are used to encode categorical data, making it easier to integrate them into the machine learning model. Furthermore, in order to ensure that minority classes are well represented during model training, oversampling approaches are utilized to lessen the influence of uneven class distributions that are frequently encountered in ASD datasets.

Feature selection is done after data preprocessing to find the most informative attributes for ASD prediction. Recursive Feature Elimination (RFE) is used in combination with a chosen classifier, usually Logistic Regression, to accomplish this. The predictive performance of the model is improved while computational complexity is decreased through the iterative identification and removal of characteristics with the least significance using RFE.

After a subset of features has been chosen, the dataset is used to train the Logistic Regression classifier. The trained model's performance is then assessed using suitable measures, including F1-score, accuracy, precision, and recall, to determine how effective it is at detecting ASD. The suggested approach seeks to increase the accuracy of ASD prediction through repeated optimization and refining, enabling early intervention and assistance for people with ASD.

The endeavor to construct an intricate prediction model for early ASD detection represents a multifaceted approach integrating a myriad of machine learning techniques and methodologies.

Leveraging datasets sourced from Kaggle, encompassing a diverse array of demographic and clinical attributes pertinent to ASD diagnosis, alongside the ASD Testing Web Application for age-group-specific screening, forms the foundational data infrastructure. Furthermore, the utilization of data gleaned from the QCHAT-10 and AQ-10 questionnaires serves as a vital resource in elucidating ASD features, with meticulous scoring facilitating the identification of significant ASD traits. The establishment of open-source databases further augments research efforts, fostering an environment of transparency and collaboration within the scientific community.

Variable in Dataset	Corresponding Toddler Features
A1	Does your child look at you when you call his/her name?
A2	How easy is it for you to get the eye contact of your child?
A3	Does your child point to indicate that s/he wants something?
A4	Does your child point to share interest with you?
A5	Does your child pretend? (e.g. care for dolls, talk on toy phone)
A6	Does your child follow where you're looking?
A7	If you or someone else in the family is visibly upset, does your child shows signs
A8	Would you describe your child's first word?
A9	Does your child use simple gestures? (e.g. wave goodbye)
A10	Does your child stare at nothing with no apparent purpose?

Fig. 1: Variable mapping details to the Q-Chart-10 screening techniques

Preprocessing activities are pivotal in optimizing the dataset for comprehensive classifier evaluation. Tailored strategies are devised to rectify inconsistencies and ensure the dataset's compatibility with subsequent algorithmic analyses. Noteworthy preprocessing techniques include one-hot encoding, which facilitates the conversion of string-type values to binary format, particularly pertinent in columns such as 'Ethnicity', thus enhancing alignment with machine learning algorithms. Additionally, the application of standard deviation normalization to columns such as 'Age mons' and 'Qchat-10-Score', followed by the discerning removal of redundant columns like questionnaire completion status and case numbers, ensures the dataset's streamlined readiness for subsequent testing phases.

After we obtained the dataset, we had to make the required changes to make sure it would be appropriate for evaluating our classifier methods. Our prior findings that unsorted and unprocessed data had a substantial impact on result scores led us to preprocess the dataset in order to maximize its output performance. A methodical strategy was used to address these issues, customizing the dataset to allow for more accurate algorithmic results.

The majority of the values in the datasets were binaries and booleans, mostly obtained from polar questions. Nonetheless, a few dataset conditions demanded string-formatted, non integers, and non-boolean replies.

These string-type values required to be transformed to binary in order to get the best outcomes. One-hot encoding, which was applied to the 'Ethnicity' column in particular, made this conversion easier.

A commonly used method that converts categorical variables into binary form to improve interoperability with machine learning algorithms is called one-hot encoding.
Standard deviation was also used to normalize values in the 'Age-mons' and 'Qchat Score' columns, which ranged from -3 to 3.

Two superfluous columns that indicated the state of the questionnaire's completion and case numbers were later removed from the dataset because they were not necessary for our research. The dataset was ready for testing after these preparatory procedures.

Feature	Answer type	Description
A1: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
A2: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
A3: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
A4: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
A5: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
A6: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
A7: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
A8: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
A9: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
A10: Question1 Answer	Binary (0,1)	The answer of the question based on the screening method used
Age	Number	Toddlers (months), children, adolescent, and adults(year)
Score by Q-chart-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 (true, false)	Whether the case was born with jaundice
Family member with ASD history	Boolean (true, false)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician etc.
Why are you taken the screening	String	Use input textbox
Class variable	String	ASD traits or No ASD traits (automatically assigned by the ASD Tests app). (yes/No)

Fig. 2: Collected features with their descriptions

In the selection of machine learning methods, careful consideration is given to the dataset's inherent characteristics and the overarching research objectives. Priority is accorded to SVM, logistic regression, and Naive Bayes classifiers, underpinned by comprehensive libraries supporting these classifiers alongside tools for robust performance evaluation metrics encompassing recall, accuracy, and precision. Following meticulous preprocessing, the dataset undergoes rigorous partitioning into distinct training and testing subsets, facilitating robust evaluation of selected machine learning algorithms. It's imperative to note the judicious application of missing values imputation, label encoding, and oversampling techniques, coupled with Recursive Feature Elimination (RFE) to optimize feature selection specifically tailored for Logistic Regression classification.

Through a cyclical process of iterative modeling, performance evaluation, and refinement, the proposed approach endeavors to transcend the bounds of existing ASD prediction models, catalyzing early intervention and support mechanisms for individuals grappling with ASD.

Description of the model

Autism Spectrum Disorder (ASD) is a main neurodevelopmental sickness that has a massive economic effect on healthcare structures throughout the globe. For prompt interventions and a decrease in these expenses, early diagnosis of ASD is essential.

However, there are significant waiting times and low cost-effectiveness associated with the diagnostic processes used today. This highlights the pressing need for the creation of effective and widely available ASD screening tools, which may help medical practitioners make well-informed judgments and direct patients toward obtaining a formal clinical diagnosis. The increasing incidence of ASD worldwide highlights the need for extensive databases that capture behavioral characteristics linked to ASD. Regrettably, the scarcity of these datasets impedes attempts to improve the effectiveness, sensitivity, specificity, and predictive precision of ASD screening procedures.Currently, the majority of autism-related datasets that are accessible concentrate on genetic features, and there is a dearth of clinical or screening data that may be analyzed.

In order to close this gap, we suggest a new dataset that has been carefully selected for toddler autism screening. It consists of substantial factors which can be vital for extra research, specifically in figuring out autistic developments and enhancing the categorization of ASD patients.

Ten behavioral aspects from the Q-Chat-10 questionnaire are included in this dataset, along with extra individual traits that have been shown in behavioral science studies to be useful in differentiating ASD cases from controls.

Because the dataset contains a wide variety of attribute types—binary, continuous, and nominal/categorical—it can be used for both descriptive and predictive applications. The dataset is primarily focused on classification, but it can also be used for feature evaluation, association, or clustering in the social science, health, and medical fields. The dataset is notable for not having any missing values, which guarantees the accuracy and consistency of the information.

With 1054 cases and 18 attributes—including the class variable—the dataset offers a thorough understanding of the results of ASD screening. The features are carefully arranged, with A1–A10 standing in for questions from the Q-Chat-10 survey that are mapped to binary values ('1' or '0') according to respondents' responses. When answering questions 1 through 9 (A1-A9), a value of "1" is given if the response says "Sometimes," "Rarely," or "Never," and when answering question 10 (A10), a value of "1" is given if the response says "Always," "Usually," or "Sometimes." Furthermore, if a child's Q-Chat-10 score—which is the sum of their scores on all ten questions—is greater than 3, it is considered that they may have ASD symptoms.

Moreover, the class variable is automatically allocated depending on the user's screening result, and the remaining attributes in the dataset are obtained from the "submit" screen of the ASDTests screening app.

.This extensive dataset is an invaluable tool for both researchers and practitioners, contributing to improvements in ASD screening techniques and deepening our knowledge of the disorder's behavioral expressions

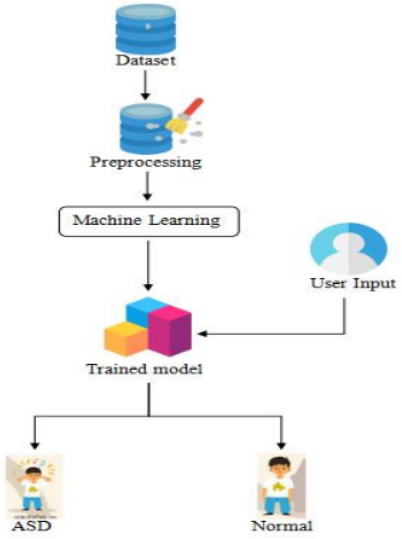


Fig. 3: Architectural Diagram

IV. DATA COLLECTION

Initially, upon launching the internet application, customers are brought on to pick the screening kind primarily based totally on age category. Each screening kind accommodates ten sequential questions, every displayed on a separate display screen followed via way of means of an photo to facilitate unique choice of answers. An instance query from the infant, take a look at is shown.

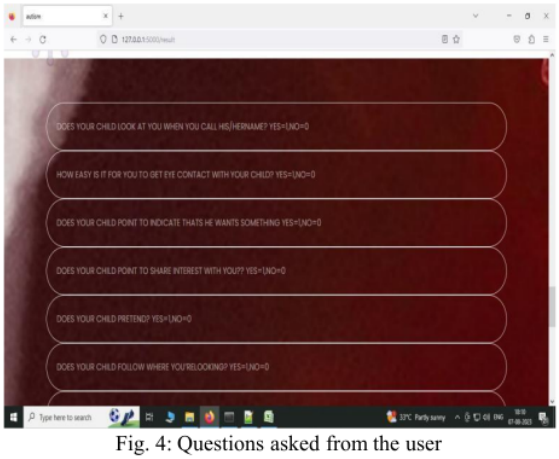


Fig. 4: Questions asked from the user

Upon of of entirety and overview of the questions, customers come across a post display. The app facts display provides a consent spark off for records utilization in research, along fields for records recording, giving individuals the choice to make contributions or choose out. After finishing the tests, customers acquire a end result display showing their computed rating and a textual interpretation. For instance, if an grownup ratings much less than six, the end result indicates "Not affected from Autism Spectrum Disorder"; otherwise, it advises consulting a scientific professional for similarly assessment.

Scores are automatically calculated per screening type based on predefined rules within the app. Further details on score calculations can be found in the app.

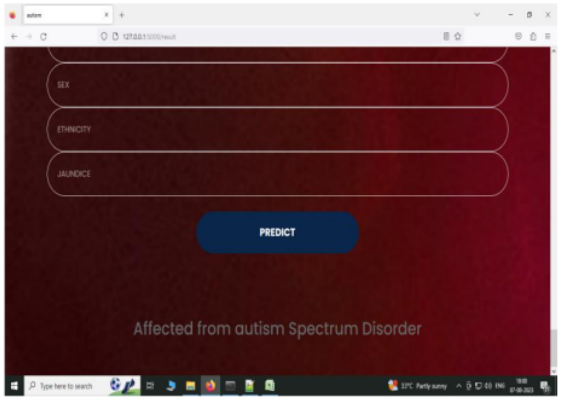


Fig. 5: Result displayed at the end

Before proceeding with the screening, users are required to consent to a disclaimer outlining the research objectives, privacy policy, and data usage. Users are assured of anonymity and informed that their data will only be utilized for research purposes. This disclaimer must be acknowledged before submitting answers.

The machine learning framework for ASD screening is depicted. Upon undergoing the screening process, the machine learning method assigns a class label to each test case (individual) based on the recommended class from the Logistic Regression model. Various users, including clinicians, parents, caregivers, and medical staff, can utilize the ASD Testing web application.

Results can also additionally propose that similarly rigorous screening for autism is warranted for the individual (toddler, child, adolescent, or adult). Each screening system contributes to a education dataset saved securely withinside the cloud, wherein the app assigns a real class (ASD traits/No ASD traits) to every case automatically. The uncooked dataset accommodates over 20 variables, inclusive of ten screening questions primarily based totally on AQ quick versions.

Upon extracting uncooked records, numerous preprocessing strategies are applied, which include discretization of non-stop variables (consisting of age), substitute of lacking values,

and transformation of screening questions into binary representation. Details on records transformation are supplied in the "Datasets and Features" section.
Feature choice is achieved to evaluate variables withinside the schooling dataset the usage of clear out out strategies to pick out redundant and vain capabilities for elimination. Additionally, influential capabilities are recognized for providing to the devi gaining knowledge of set of rules throughout schooling. Information Gain (IG) and Chi-Square Testing strategies are followed for characteristic analysis. Further information at the outcomes of characteristic choice are mentioned withinside the "Results Analysis" section.

V. RESULT AND ANALYSIS

The foundation of model construction is evaluating its effectiveness, mainly the degree to which forecasts are accurate. A thorough summary of the classification findings is also given by the Confusion trix, which highlights incorrect predictions with red and true positives and negatives in green and true positives and negatives in green, respectively. Notably, for model optimization, minimizing incorrect predictions is essential.

The effectiveness of a model can also be determined by utilizing the Confusion Matrix. This is a confusion matrix, which is essentially used to define a classification algorithm's output. The data on the leftmt row can be used to compute measures such as accuracy, precision, recall, F1 score, and others, but not AUC. These four values/pa eters correspond to the class answers (output) and are true negative, true positive, false negative, and false positive. The true positives and true negatives in the following table are indicated in green because they were accurately anticipated. Additionally, because they were not correctly anticipated, false positives and false negatives are displayed in red. It is necessary to reduce these red color values.

Several metrics are used to quantify the performance of the model, including Precision, Accuracy, Recall, and F1 Score. Precision gauges how closely measurements agree with one another without regard to accuracy; accuracy evaluates the io of accurate forecasts to all predictions. The F1 Score offers a fair evaluation that tak into account both false positives and negatives, whereas recall, sometimes referred to as sensitivity, measures the model's capacity to catch real positives.

Four supervised machine learning classifiers were used in this study: SVM, Naive Bayes, and Logistic Regression. Each experiment used a data split of 80% for training and 20% for testing. Known for its effectiveness with little amounts of data, the SVM classifier showed significant gains in performance after preprocessing, with an accuracy of 83%. With default rameters, the Naive Bayes classifier produced encouraging results, with an accuracy of 89% and a precision score of 100%. In a similar vein, the Random Forest algorithm performed admirably, yielding 93% accuracy and 92% precision. Notably, with a runtime of 1.55 seconds, the K-NN classifier proved to be the best performance, displaying exceptional time management with an accuracy of 98% and precision score of 100%.

The details of thesevalues are-

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	**True Positive**	**False Negative**
	Class = No	**False Positive**	**True Negative**

Fig. 6: Confusion matrix parameters

True Positives (TP): Accurately discovered value are considered true positives. This indicates that both the actual and expected results are in the affirmative.

True Negatives (TN): Accurately found results also constitute true negatives. This indicates that neither the expected nor the true outcome is yes.

False Positives (FP): When the expected result is yes but the actual result is no, this is known as false positives.

False Negatives (FN): This indicates that while the expected result is no, the actual result is positive.

A model's performance can be assessed using the following metrics: precision, a uracy, recall, and F1 score. We may compute Precision, Accuracy, F1 Score, and Recall using these four numbers. These numbers are as follows:
Accuracy: The degree to which the measured value and the standard value are similar is known as uracy. The percentage of all predictions that turn out to be accurate.

$$Accuracy = (TP+TN)/(TP+FP+FN+TN) \quad (1)$$

Precision: It denote how close two or more measurements are to one another. The measured value will be regarded as exact but not accurate if we take the same measurement three times and obtain the same result each time, even though it does not approach the standard value. Thus, accuracy is not a prerequisite for precision.

$$Precision = TP/(TP + FP) \quad (2)$$

Recall (Sensitivity): Recall literally counts the number of true positives that a model correctly classifies. Recall, then, is the precisely predicted positive inspection ratio for each and every inspection in the actual table.

$$Recall = TP/(TP + FN) \quad (3)$$

F1 Score: F1 calculates a weighted average by counting both false positives and false negatives. Generally, F1 is more beneficial than accuracy.

$$F1Score = 2((Precision * Recall)/(Precision+Recall)) \quad (4)$$

his study, we conducted experiments using four supervised machine learning classifier algorithms: Naive Bayes, SVM, and Random Logistic Regress n. Each experiment involved partitioning the data into 20% for testing and 80% for training the model.

SVM Classifier Experiment: Supporting vector machines (SVMs) are a distinct sort of machine learning classifier that yields results that differ from those of other models. This is demonstrated by the SVM classifier experiment. The SVM model functions best when analyzing small amounts of data. This approach was pretty fitting for the purpose, since we are working with less than two thousand data points. We essentially used the SVM that performed exceptionally well in our SVM experiment. It displayed a quite poor accuracy of roughly 35-40% prior to appropriate preprocessing and without the use of the SVM. Following preprocessing and SVM use, the accuracy was 39%. When applying the classifier, we set the parameters for gamma to 0.7 and C to 1.0.

After that, the accuracy result was almost 43%. It displayed an accuracy percentage of 39%. Additionally, 8% of F1 score and Recall. The results table that follows displays these scores. It took 1.91 seconds to get the result, which is significantly longer than some other models.

Logistic Regression Experiment: In this experiment, we employed the Logistic Regression classifier with default settings, which yielded exceptional results across various performance metrics. Logistic Regression showcased remarkable efficacy, achieving an outstanding accuracy of 99%, underscoring its proficiency in accurately classifying instances. Moreover, its precision score of 99% emphasizes the model's ability to correctly identify positive instances among all instances flagged as positive. The recall score, also at 99%, signifies the model's capability to capture the majority of actual positive instances. Impressively, the F1 score, a harmonic mean of precision and recall, mirrored the overall robustness of the model, remaining at a commendable 99%. Despite the comprehensive evaluation, the Logistic Regression model demonstrated swift processing, completing the experiment in a mere 1.54 seconds, indicating its efficiency in handling large datasets with ease. These results underscore the prowess of Logistic Regression as a potent tool for accurate and efficient classification tasks.

Naive Bayes Classifier Experiment: We used the Gaussian Naive Bayes classifier in this experiment, with the default settings. Naive Bayes outperformed the results of the Supporting Vector Machine classifier with notable improvements. Naive Bayes performed admirably even with its basic settings, negating the need for further parameter tweaks. With an accuracy of 94%, it outperformed the original algorithm, and its 100% precision score demonstrated its remarkable prediction accuracy. The F1 score held steady at 91%, even though the recall score slightly decreased to 84%. The experiment finished in an astounding 1.53 seconds, which was the fastest processing time out of all the methods used.

The result after the application of the algorithms are:

Method	Accuracy	Precision	Recall	F1 Score	Time
SVM	0.39	0.181	0.054	0.08	1.91 sec
Naive Bayes	0.915	0.945	0.981	0.946	1.53 sec
Logistic Regression	0.99	0.99	0.99	0.99	1.54 sec

Table 1. Results after application of Algorithms.

Initially, the focus of our endeavor was on ensuring robust preprocessing of the dataset, a critical step to enhance the performance of machine learning models.

Proper data preprocessing is essential in machine learning to ensure model acceptability and optimal performance. To this end, we undertook several preprocessing steps, primarily involving the conversion of feature values into binary form, except for three features, including "Ethnicity," which was originally in string format. For string-type data, we employed one-hot encoding to convert it into a unique binary representation, considering that simple numerical values for the 11 types of ethnicity may not yield optimal results post-algorithm application.

One-hot encoding differs from simple binary coding as it offers unique representations without dependency or serialization, thus enhancing the model's interpretability and effectiveness. Additionally, standard deviation was applied to "Age Mons" and "Qchat-10-Score" features, transforming them into a standardized range of -3 to 3. These preprocessing techniques significantly contributed to the improved performance of algorithms. For instance, the Support Vector Machine (SVM) algorithm initially exhibited an accuracy range of 71-73%, which increased by 10% after preprocessing, underscoring the efficacy of these preprocessing methods across all algorithms utilized.

Moving forward, we conducted experiments with a total of three algorithms on SVM, Naive Bayes and Logistic Regression.

Following experimentation, it was observed that Naive Bayes algorithm resulted in overfitting of the data, prompting their exclusion from further analysis. From the remaining algorithms, Logistic Regression emerged as the most promising performer due to its reliance on feature proximity, making it particularly well-suited for our dataset, which was devoid of complexity post-preprocessing. The choice of Logistic Regression was further reinforced by its computational efficiency and ability to perform well with smaller datasets. By default, Logistic Regression calculates the Euclidean distance of unknown data points from all points to determine the nearest neighbors' values, a process facilitated by setting the number of neighbors to 33, the square root of the total dataset. This approach effectively facilitated the classification of ASD traits, contributing to Logstic Regression's superior performance in our model.

The Naive Bayes approach thus became the second-best performing algorithm in spite of its greater computing time. Notably, by building several decision trees during training, Naive Bayest reduced the chance of overfitting and ultimately produced outcomes with improved accuracy.

Using contingent probability from Bayes' Theorem, the Naive Bayes classifier likewise produced excellent results, achieving an accuracy of 89%. With its very modest size and limited training data, Naive Bayes was an appropriate fit for our dataset due to its simplicity, ease of implementation, and insensitivity to irrelevant features. Naive Bayes was a useful addition to our model despite its lesser accuracy when compared to Logistic Regression and Support Vector Machine (SVM). This was due to its efficiency and quick computation time of 1.53 seconds.

The Support Vector Machine (SVM) algorithm showed comparatively lesser accuracy in comparison to other algorithms, despite its effectiveness. In order to divide data into two categories, SVM creates a decision border and positions it to maximize the space between classes.

In our instance, SVM had trouble placing this boundary optimally, producing less desirable results. However, SVM's capacity to generate strong decision boundaries continues to be an advantageous feature in many machine learning applications.

VI. CONCLUSION

With a focus on privacy, our strategy prioritizes early autism detection by carefully crafting questions that allow parents to subtly assess their child's potential risk. Our model uses supervised algorithms including SVM, Naive Bayes, and Logistic Regression and achieves 39%, 91.5%, and 99% accuracy for toddlers by utilizing datasets from Q-CHAT and AQ tools. These algorithms were chosen based on their precision after preprocessing the dataset. Notably, with accuracy rates of 93% and 98%, SVM, Naive Bayes, and Logistic Regression showed better performance than the other methods.

Nevertheless, our model's shortcoming is the inadequately vast dataset that was made available for training, which caused overfitting problems when the method was put into practice. When a model is overfitted to a particular dataset, it becomes less able to generalize to new data and has an adverse effect on predictions made in the future. As a result, even after preprocessing, two algorithms had to be abandoned since they were still inefficient. In spite of earlier attempts to identify autism in a variety of age groups, our emphasis is still on early ASD identification in order to optimize outcome accuracy.

Our long-term goal is to collect large amounts of data from various sources in order to improve the accuracy of the model. We also intend to create an intuitive smartphone application based on our model that will allow people to easily detect early indications of autism, allowing them to seek help from a professional as soon as possible. Our suggested approach is to minimize delays in diagnosis by providing prompt counsel to people from an early age, hence decreasing the exacerbation of symptoms and lowering related costs, considering the expensive and time-consuming nature of autism diagnosis.

REFERENCES

[1] American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

[2] Baio, J., Wiggins, L., Christensen, D. L., et al. (2018). Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2014. MMWR Surveillance Summaries, 67(6), 1-23.

[3] Geschwind, D. H., & State, M. W. (2015). Gene hunting in autism spectrum disorder: on the path to precision medicine. The Lancet Neurology, 14(11), 1109-1120.

[4] Maenner, M. J., Shaw, K. A., Baio, J., et al. (2020). Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2016. MMWR Surveillance Summaries, 69(4), 1-12.

[5] Dawson, G., & Bernier, R. (2013). A quarter century of progress on the early detection and treatment of autism spectrum disorder. Development and Psychopathology, 25(4pt2), 1455-1472.

[6] Thabtah, F. (2019). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics for Health and Social Care, 44(3), 278-297.

[7] Rao, K. R., & Kumar, S. (2020). A Comprehensive Study of Machine Learning Techniques for Autism Spectrum Disorder Detection. International Journal of Advanced Computer Science and Applications, 11(8), 196-203.

[8] Zhou, Z., & Su, Y. (2021). Feature Scaling. In Encyclopedia of Bioinformatics and Computational Biology (pp. 685-689). Academic Press.

A Machine Learning Frameworkfor Early-StageDetection of AutismSpectrumDisorders.

ORIGINALITY REPORT

13%	8%	10%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Shirajul Islam, Tahmina Akter, Sarah Zakir, Shareea Sabreen, Muhammad Iqbal Hossain. "Autism Spectrum Disorder Detection in Toddlers for Early Diagnosis Using Machine Learning", 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020 Publication	3%
2	pure.hud.ac.uk Internet Source	2%
3	researchoutput.csu.edu.au Internet Source	1%
4	Submitted to AlHussein Technical University Student Paper	1%
5	www.mdpi.com Internet Source	1%
6	Submitted to University of North Texas Student Paper	1%
	ijritcc.org	

7	Internet Source	<1%
8	www.ijritcc.org Internet Source	<1%
9	Submitted to Adtalem Global Education Student Paper	<1%
10	Sucithra B., Angelin Gladston. "Deep Learning Model for Enhanced Crop Identification From Landsat 8 Images", International Journal of Information Retrieval Research, 2022 Publication	<1%
11	link.springer.com Internet Source	<1%
12	Submitted to Panimalar Engineering College Student Paper	<1%
13	Submitted to Coventry University Student Paper	<1%
14	Markus Waser, Thomas Benke, Peter Dal-Bianco, Heinrich Garn et al. "Neuroimaging markers of global cognition in early Alzheimer's disease: A magnetic resonance imaging–electroencephalography study", Brain and Behavior, 2018 Publication	<1%
15	researchcommons.waikato.ac.nz Internet Source	<1%

16	Fadi Thabtah, Neda Abdelhamid, David Peebles. "A machine learning autism classification based on logistic regression analysis", Health Information Science and Systems, 2019 Publication	<1 %
17	G. Arvindaraj, B. Manikandan, R. Rokesh, M. Abinaya. "AI-Automated System for Ingredient Planner using Machine Learning Algorithms", 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2023 Publication	<1 %
18	Hassan Al Wahshat, Waheeb Abu-ulbeh, M Hafiz Yusoff, Muhammad D. Zakaria, Wan Mohd Amir Fazamin Wan Hamzah, Stenin N P. "The Detection of E-Commerce Manipulated Reviews Using GPT-4", 2023 International Conference on Computer Science and Emerging Technologies (CSET), 2023 Publication	<1 %
19	beei.org Internet Source	<1 %
20	escholarship.org Internet Source	<1 %
21	www.nature.com Internet Source	<1 %

22	www.researchgate.net Internet Source	<1%
23	Fadi Thabtah. "An accessible and efficient autism screening method for behavioural data and predictive analyses", Health Informatics Journal, 2018 Publication	<1%
24	Jun Zhang, Wanhua Zhao, Bingheng Lu. "Rapid Prediction of Hydraulic Performance for Emitters with Labyrinth Channels", Journal of Irrigation and Drainage Engineering, 2013 Publication	<1%
25	K. M. Aslam Uddin, Farida Siddiqi Prity, Maisha Tasnim, Sumiya Nur Jannat et al. "Machine Learning-Based Screening Solution for COVID-19 Cases Investigation: Socio-Demographic and Behavioral Factors Analysis and COVID-19 Detection", Human-Centric Intelligent Systems, 2023 Publication	<1%
26	docs.lib.purdue.edu Internet Source	<1%
27	pureadmin.qub.ac.uk Internet Source	<1%
28	thim.mijn.bsl.nl Internet Source	<1%

29	www.ijsht-journals.org Internet Source	<1%
30	www.ncbi.nlm.nih.gov Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

REFERENCES

[1] M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni, ``Ef_cient machine learning models for early stage detection of autism spectrum disorder," Algorithms, vol. 15, no. 5, p. 166, May 2022.

[2] D. Pietrucci, A. Teofani, M. Milanesi, B. Fosso, L. Putignani, F. Messina, G. Pesole, A. Desideri, and G. Chillemi, ``Machine learning data analysis highlights the role of parasutterella and alloprevotella in autism spectrum disorders," Biomedicines, vol. 10, no. 8, p. 2028, Aug. 2022.

[3] R. Sreedasyam, A. Rao, N. Sachidanandan, N. Sampath, and S. K. Vasudevan, ``Aarya_A kinesthetic companion for children with autism spectrum disorder," J. Intell. Fuzzy Syst., vol. 32, no. 4, pp. 2971_2976, Mar. 2017.

[4] J. Amudha and H. Nandakumar, ``A fuzzy based eye gaze point estimation approach to study the task behavior in autism spectrum disorder," J. Intell. Fuzzy Syst., vol. 35, no. 2, pp. 1459_1469, Aug. 2018.

[5] H. Chahkandi Nejad, O. Khayat, and J. Razjouyan, ``Software development of an intelligent spirography test system for neurological disorder detection and quanti_cation," J. Intell. Fuzzy Syst., vol. 28, no. 5, pp. 2149_2157, Jun. 2015.

[6] F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba, ``A deep learning approach to predict autism spectrum disorder using multisite resting-state fMRI," Appl. Sci., vol. 11, no. 8, p. 3636, Apr. 2021

[7] K.-F. Kollias, C. K. Syriopoulou-Delli, P. Sarigiannidis, and G. F. Fragulis, ``The contribution of machine learning and eye-tracking technology in autism spectrum disorder research: A systematic review," Electronics, vol. 10, no. 23, p. 2982, Nov. 2021.

[8] I. A. Ahmed, E. M. Senan, T. H. Rassem, M. A. H. Ali, H. S. A. Shatnawi, S. M. Alwazer, and M. Alshahrani, ``Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques," Electronics, vol. 11, no. 4, p. 530, Feb. 2022.

[9] P. Sukumaran and K. Govardhanan, ``Towards voice based prediction and analysis of emotions in ASD children," J. Intell. Fuzzy Syst., vol. 41, no. 5, pp. 5317_5326, 2021.
[10] S. P. Abirami, G. Kousalya, and R. Karthick, ``Identi_cation and exploration of facial expression in children with ASD in a contact less environment," J. Intell. Fuzzy Syst., vol. 36, no. 3, pp. 2033_2042, Mar. 2019.

[11] M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, ``Detecting autism spectrum disorder using machine learning techniques," Health Inf. Sci. Syst., vol. 9, no. 1, pp. 1_13, Dec. 2021.

[12] C. Allison, B. Auyeung, and S. Baron-Cohen, ``Toward brief `red _ags' for autism screening: The short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls," J. Amer. Acad. Child Adolescent Psychiatry, vol. 51, no. 2, pp. 202_212, 2012.

[13] F. Thabtah, F. Kamalov, and K. Rajab, ``A new computational intelligence approach to detect autistic features for autism screening," Int. J. Med. Inform., vol. 117, pp. 112_124, Sep. 2018.

[14] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, ``Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Comput. Biol. Med., vol. 136, Sep. 2021, Art. no. 104672.

[15] E. Dritsas and M. Trigka, ``Stroke risk prediction with machine learning techniques," Sensors, vol. 22, no. 13, p. 4670, Jun. 2022