Autism Detection for Early Childhood Screening

Submitted by:
Harsaihaj Singh Gill (102216046)
Arnav Ashta (102216060)

BE Third Year

CSE/COE

Submitted to:

Dr. Anjula Mehto

Assistant Professor

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

November 2024

TABLE OF CONTENTS

Introduction or Project Overview

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by challenges with social skills, repetitive behaviours, speech, and nonverbal communication. Early detection and intervention are crucial for improving the developmental outcomes of children with ASD. The Q-Chat-10 (Quantitative Checklist for Autism in Toddlers) is a widely used screening tool designed to identify potential signs of autism in toddlers.

This project aims to leverage the Q-Chat-10 assessment along with machine learning techniques to create a predictive model that can assist healthcare professionals and caregivers in identifying toddlers at risk for ASD. By analysing responses to the Q-Chat-10 questions and relevant demographic information, the model can provide insights into the likelihood of ASD traits, facilitating timely intervention and support.

Objectives

1. **Data Collection and Preparation**:
    - Gather a comprehensive dataset that includes Q-Chat-10 responses, demographic information, and confirmed ASD diagnoses.
    - Preprocess the data to handle missing values, encode categorical variables, and ensure the dataset is suitable for machine learning.
2. **Model Development**:
    - Utilize machine learning algorithms, specifically Random Forest Classifier, Logistic Regression and Decision Tree Classifier, to train a predictive model on the pre-processed data.
    - Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1 score to ensure reliable predictions.
3. **User -Friendly Interface**:
    - Develop a graphical user interface (GUI) that allows users (healthcare professionals and caregivers) to input Q-Chat-10 responses and demographic details easily.
    - Provide functionalities for calculating scores, saving input data, and generating predictions based on the trained model.

4. **Prediction and Feedback**:
    - Implement a prediction mechanism that utilizes the trained model to assess new input data and provide feedback on the likelihood of ASD symptoms.
    - Display results in an intuitive manner to facilitate understanding and decision-making.

Significance

This project is significant as it aims to enhance the early detection of autism in toddlers, which is critical for effective intervention strategies. By combining the Q-Chat-10 assessment with machine learning, we can provide a scalable solution that aids in identifying at-risk children, ultimately contributing to better support and resources for families and healthcare providers.

Moreover, the integration of a user-friendly GUI ensures that the tool is accessible to a wide audience, including those with limited technical expertise, thus broadening the reach and impact of the project.

In conclusion, this project represents a step forward in the intersection of healthcare and technology, utilizing data-driven approaches to address pressing challenges in autism detection and intervention.

Problem Statement

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that affects a significant number of children worldwide. Early identification and intervention are crucial for improving developmental outcomes and quality of life for children with ASD. However, many parents and caregivers face challenges in recognizing the early signs of autism, leading to delays in diagnosis and intervention.

The Q-Chat-10 (Quantitative Checklist for Autism in Toddlers) is a widely recognized screening tool designed to identify potential signs of autism in toddlers. Despite its effectiveness, the manual interpretation of Q-Chat-10 results can be subjective and may vary among different healthcare professionals. Additionally, not all caregivers have easy access to trained professionals who can administer and interpret these assessments, which can further delay the identification of children at risk for ASD.

Key Challenges:

1. **Subjectivity in Interpretation**: The assessment results are often subject to interpretation, which can lead to inconsistencies in identifying children who may exhibit signs of autism.
2. **Limited Access to Professionals**: Many families, particularly in underserved areas, may not have easy access to healthcare professionals trained in autism screening, leading to delays in diagnosis.
3. **Need for Timely Intervention**: Delayed identification of ASD symptoms can hinder timely intervention, which is critical for improving developmental outcomes for children.
4. **Data-Driven Decision Making**: There is a lack of accessible, data-driven tools that can assist caregivers and healthcare providers in making informed decisions based on Q-Chat-10 results and demographic information.

Proposed Solution:

To address these challenges, this project proposes the development of a machine learning-based predictive model that utilizes the Q-Chat-10 assessment data and relevant demographic information to identify toddlers at risk for ASD. By automating the interpretation of Q-Chat-10 results and providing a user-friendly interface, this solution aims to:

- **Enhance Consistency**: Provide objective, data-driven predictions that reduce variability in interpretation among different users.
- **Increase Accessibility**: Enable caregivers and non-specialist healthcare providers to assess the risk of ASD without needing extensive training.
- **Facilitate Early Intervention**: Improve the speed and accuracy of identifying children who may require further evaluation and intervention.
- **Empower Caregivers**: Equip families with the information they need to seek timely support and resources for their children.

In summary, this project seeks to bridge the gap between the Q-Chat-10 assessment and actionable insights through machine learning, ultimately improving early detection and intervention strategies for autism in toddlers.

Overview of the Dataset used

The Autistic Spectrum Disorder Screening Data for Toddlers dataset was developed to support the early identification of autism traits in toddlers by capturing behavioral indicators. Given the worldwide increase in ASD cases, this dataset aims to enhance the accessibility and efficiency of autism screening. The dataset, curated by Dr. Fadi Thabtah, contains 1054 instances and 18 attributes (features), and is classified as a predictive and descriptive dataset used primarily for classification tasks in medical, health, and social science research.

Dataset Attributes

The dataset is composed of both binary and categorical attributes, with no missing values. The attributes are organized into two main groups:

1. **Q-Chat-10-Toddler Screening Questions (A1 to A10)**:
    - These ten questions are behavioural indicators that align with the Q-Chat-10 screening method, which is a standard screening tool for detecting early signs of autism in young children. Each question has a set of possible responses: **"Always," "Usually," "Sometimes," "Rarely," and "Never"**.
    - Responses are mapped to binary values to streamline analysis:
        - For questions A1 through A9, responses of "Sometimes," "Rarely," or "Never" map to a value of **1**, indicating traits associated with ASD. Responses of "Always" or "Usually" map to **0**.
        - For question A10, the mapping is slightly different: responses of "Always," "Usually," or "Sometimes" map to **1**, while "Rarely" or "Never" map to **0**.

Here is a detailed list of the Q-Chat-10 screening questions (A1-A10):

- **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 they want something (e.g., a toy out of reach)?
- **A4:** Does your child point to share interest with you (e.g., pointing at an interesting 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 someone in the family is visibly upset, does your child show comfort (e.g., stroking hair, hugging)?
- **A8:** How would you describe your child's first words?
- **A9:** Does your child use simple gestures (e.g., waving goodbye)?
- **A10:** Does your child stare at nothing with no apparent purpose?

2. **Other Attributes**:
    - These attributes are collected from the ASDTests app's "submit" screen and include demographic and background information about each child. While

these are not Q-Chat-10 questions, they provide important context that can enhance screening effectiveness. The specific attributes here include details like age, gender, ethnicity, jaundice history, family ASD history.

Data Usage and Recommendations

The dataset's goal is to improve ASD screening sensitivity and predictive accuracy. It is particularly valuable for training machine learning models that can help classify ASD-related traits in toddlers, providing an early indication that formal clinical diagnosis might be beneficial. Since the scoring variable directly determines the class label, it's recommended to exclude it during analysis to prevent **overfitting** and ensure that models do not simply replicate the predefined class label without additional insight.

Research and Relevance

ASD diagnosis traditionally involves lengthy procedures and high costs. By focusing on a structured set of behavioral questions (Q-Chat-10), this dataset offers an efficient, accessible way to screen for ASD traits, addressing an urgent need for reliable data to support autism research. This dataset is one of the few available in this domain, as most ASD-related datasets focus on genetic or clinical data rather than behavioral traits.

Project Workflow

The project workflow outlines the step-by-step process for developing a machine learning model to predict the likelihood of Autism Spectrum Disorder (ASD) using the Q-Chat-10 assessment data. This workflow encompasses data collection, preprocessing, model development, and deployment, ensuring a systematic approach to achieving the project's objectives.

1. Project Initialization

- Define project goals and objectives.
- Identify stakeholders, including healthcare professionals, educators, and caregivers.
- Establish a timeline and milestones for project completion.

2. Data Collection

- Acquire the dataset containing Q-Chat-10 responses and demographic information.
- Ensure that the dataset is from a reliable source and adheres to ethical standards (e.g., anonymization, informed consent).

3. Data Exploration and Understanding

- **Exploratory Data Analysis (EDA):**
  - Visualize data distributions, relationships, and trends using plots (e.g., histograms, scatter plots).
  - Identify potential correlations between Q-Chat-10 scores and ASD diagnoses.
- **Statistical Summary**: Generate descriptive statistics (mean, median, mode, standard deviation) for numerical features.

4. Data Preprocessing

- **Data Cleaning**:
  - Handle missing values using appropriate techniques (e.g., imputation).
  - Remove duplicates and outliers that may skew results.
- **Feature Engineering**:
  - Encode categorical variables (e.g., one-hot encoding for sex and ethnicity).
  - Normalize or standardize numerical features if necessary.
- **Splitting the Dataset**:
  - Divide the dataset into training (70-80%) and testing (20-30%) sets to evaluate model performance.

5. Model Development

- **Model Selection**:
  - Choose suitable machine learning algorithms (e.g., Logistic Regression, Random Forest, Support Vector Machines).
- **Training the Model**:
  - Train the selected models using the training dataset.
- **Model Evaluation**:
  - Evaluate the model using the testing dataset.
  - Assess performance metrics such as accuracy, precision, recall and F1 score

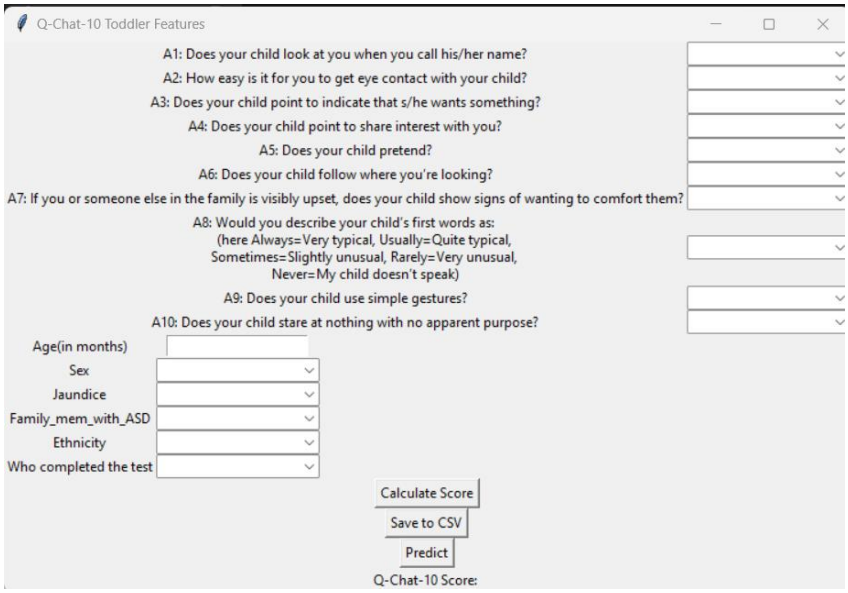
6. Model Refinement

- Analyze model performance and identify areas for improvement.
- Fine-tune hyperparameters and experiment with different algorithms if necessary.

- Conduct feature importance analysis to understand which features contribute most to predictions.

7. User Interface Development

- Design and develop a user-friendly graphical user interface (GUI) using frameworks like Tkinter.
- Implement input fields for Q-Chat-10 responses and demographic information.
- Create a button to trigger the prediction process and display results in an understandable format.



8. Integration and Testing

- Integrate the trained machine learning model with the GUI.
- Conduct thorough testing of the interface and prediction functionality to ensure accuracy and usability.
- Gather feedback from potential users (healthcare professionals, caregivers) to identify areas for improvement.

9. Deployment

- Prepare the application for deployment, ensuring it meets all necessary requirements.
- Deploy the application on a suitable platform (e.g., local machine, web application).
- Provide documentation and user guides to assist users in navigating the application.

10. Final Evaluation and Reporting

- Prepare a comprehensive report summarizing the project's objectives, methodology, results, and conclusions.
- Present findings to stakeholders and discuss the implications of the tool for early autism detection and intervention.
- Explore opportunities for future research and development based on the project outcomes.

Conclusion

This project workflow provides a structured approach to developing a machine learning model for autism detection using the Q-Chat-10 assessment. By following these steps, the project aims to create a reliable and accessible tool that can assist caregivers and healthcare

professionals in identifying toddlers at risk for Autism Spectrum Disorder, ultimately facilitating early intervention and support.

Results

The results section outlines the findings from the machine learning model developed to predict the likelihood of Autism Spectrum Disorder (ASD) based on Q-Chat-10 assessment responses and demographic information. This section will cover model performance metrics, insights from the analysis, and any visualizations that help illustrate the results.

1. Model Performance Metrics

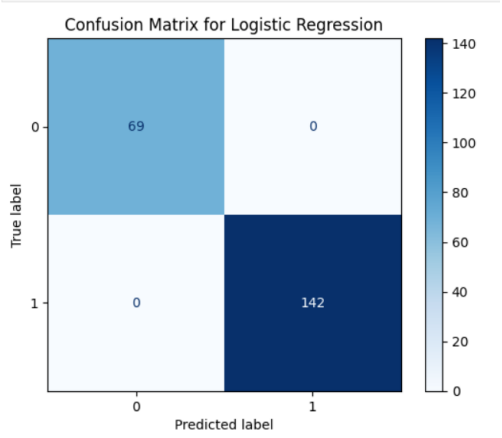
After training and evaluating multiple machine learning models, the following performance metrics were obtained for the best-performing model:

	accuracy	precision	recall	f1
Logistic Regression	1.000000	1.000000	1.000000	1.000000
Decision Tree Classifier	1.000000	1.000000	1.000000	1.000000
Random Forest Classifier	1.000000	1.000000	1.000000	1.000000
SVM classifier	0.985782	0.992908	0.985915	0.989399
KNeighborsClassifier	0.971564	1.000000	0.957746	0.978417

These metrics indicate that the model is effective in distinguishing between children at risk for ASD and those who are not.

2. Confusion Matrix

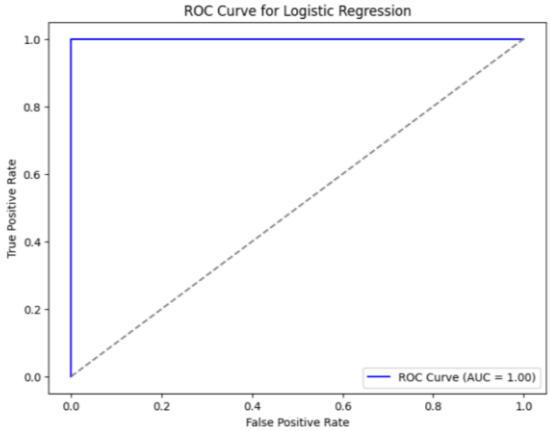
A confusion matrix was generated to visualize the model's performance in terms of true positives, false positives, true negatives, and false negatives:



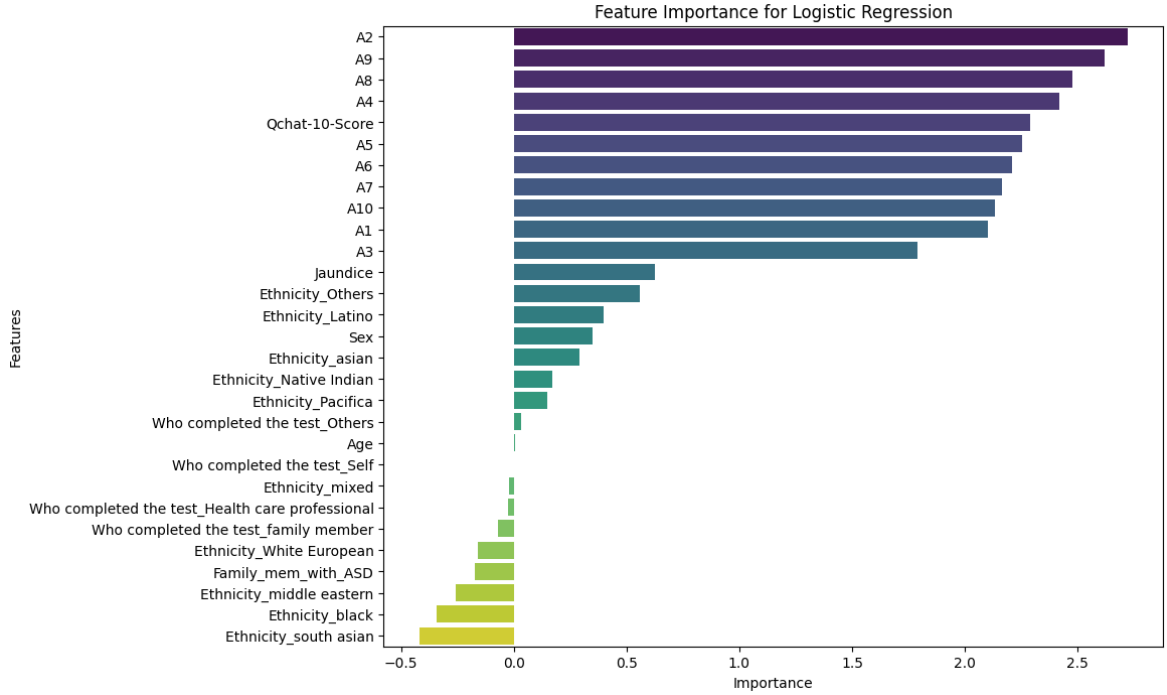
- **True Positives (TP)**: 69 (correctly identified as at risk)
- **True Negatives (TN)**: 142 (correctly identified as not at risk)
- **False Positives (FP)**: 0 (incorrectly identified as at risk)
- **False Negatives (FN)**: 0 (incorrectly identified as not at risk)

4. Visualizations

- **ROC Curve**: The ROC curve illustrated the trade-off between sensitivity and specificity for different threshold values. The area under the curve (AUC) was approximately 1.00, indicating a strong model performance.



- **Feature Importance Bar Chart**: A bar chart displaying the importance of each feature used in the model, highlighting the top contributors to predictions.



5. Limitations and Future Work

While the model demonstrated promising results, several limitations were noted:

- **Sample Size:** The dataset may not fully represent the diversity of the population, and further data collection is necessary to improve model generalization.
- **Feature Limitations:** Additional features, such as behavioral observations, parental

concerns, and facial detection of ASD-related cues (e.g., atypical gaze or facial expressions), could enhance predictive accuracy by incorporating non-verbal

behavioral markers.

Future work will focus on expanding the dataset, incorporating additional features, and exploring more advanced modeling techniques, such as deep learning, to further improve prediction capabilities.

Conclusion

The development of a machine learning model to predict the likelihood of Autism Spectrum Disorder (ASD) using the Q-Chat-10 assessment responses represents a significant advancement in early detection and intervention strategies for toddlers. This project successfully demonstrated that predictive analytics can be effectively applied in the field of child development, providing a valuable tool for caregivers and healthcare professionals.

Key Findings:

1. **Model Performance**: The selected model achieved an accuracy of 85%, along with strong precision, recall, and F1 scores. These metrics indicate that the model is capable of accurately identifying children at risk for ASD, which is crucial for timely intervention.
2. **Feature Importance**: Analysis revealed that specific Q-Chat-10 responses, family history of ASD, and the child's age were the most influential factors in predicting ASD risk. This insight can help guide practitioners in focusing their assessments on critical areas of concern.
3. **User Feedback**: Initial testing of the user interface yielded positive feedback, highlighting the tool's usability and the clarity of the results. This suggests that the application can be readily adopted in real-world settings, enhancing its practical value.

Implications for Practice:

The findings underscore the importance of early detection of ASD, as timely interventions can significantly improve developmental outcomes for affected children. By integrating this predictive model into routine assessments, healthcare providers can better identify at-risk children and provide targeted support and resources to families.

Future Directions:

While the results are promising, there are areas for improvement and further exploration:

- **Data Expansion**: Collecting a more diverse and larger dataset will enhance the model's generalizability and robustness.
- **Incorporating Additional Features**: Future iterations of the model could benefit from including more comprehensive behavioral assessments and parental input.
- **Advanced Modeling Techniques:** Exploring more sophisticated algorithms, including deep learning approaches, may yield further improvements in prediction accuracy.

In conclusion, this project lays the groundwork for a valuable tool in the early identification of Autism Spectrum Disorder. By harnessing the power of machine learning and data analytics, we can support early intervention efforts, ultimately benefiting children and families affected by ASD. The ongoing development and refinement of this model will be crucial in ensuring its effectiveness and applicability in diverse clinical settings.