

Applied Data Science - Guided Project

Detection Of Autistic Spectrum Disorder: Classification

Submitted by:

Team No.: 378

Team Lead: Gayathri Reddy Patlolla

Reg No.: 20BCI0235

Email: gayathrireddy.patlolla2020@vitstudent.ac.in

Contact no.: 7680088355

Team Member 1: Aryaa Reddy Bheemreddy

Reg No.: 20BBS0186

Email: aryaareddy.bheemreddy2020@vitstudent.ac.in

Contact no.: 817977302

Team Member 2: Jayashree Murugesan

Reg No.: 20BCE2161

Email: jayashree.murugesan2020@vitstudent.ac.in

Contact no.: 9363485553

Team Member 3: Bhargavi Gupta

Reg No.: 20BBS0172

Email: bhargavi.gupta2020@vitstudent.ac.in

Contact no.: 9999490198

Demo Video Link: <https://drive.google.com/file/d/1InvSVVA-OTVpI7t9PNnx1h7-WqA5wXx9/view?usp=sharing>

1 INTRODUCTION

1.1 Overview

The project aims to develop a machine learning model for the detection of Autism Spectrum Disorder (ASD). Complex neurodevelopmental illness known as autism is characterised by issues with speech, social interaction, and repetitive behaviours. Early ASD discovery is essential for prompt intervention and improved outcomes for those living with the illness.

To achieve this goal, the project leverages an existing dataset that encompasses a range of pertinent information, including behavioural observations, survey responses, and demographic data. This dataset serves as a valuable resource for training and testing the machine learning algorithms employed in the project. By analyzing this dataset, the algorithms can uncover underlying patterns and trends that can aid in the identification of ASD cases.

Using feature selection techniques is one of the project's essential components. These methods are used to find the dataset's elements that are most essential for accurately identifying ASD. The machine learning model can improve its capacity to accurately distinguish between people with ASD and those without the disorder by identifying the most informative traits.

To evaluate the performance of the generated machine learning models, a variety of performance indicators are utilized. These indicators include metrics such as accuracy, precision, recall, and the F1 score. Accuracy assesses the overall correctness of the model's predictions, while precision measures the model's ability to correctly identify individuals with ASD. Recall, on the other hand, evaluates the model's capability to identify all the individuals with ASD correctly. The F1 score provides a balanced assessment of precision and recall, taking both into account.

These evaluations help refine and improve the model over time, ensuring its reliability and efficiency in real-world applications for ASD detection.

1.2 Purpose

The project at hand encompasses a number of remarkable applications and accomplishments that merit attention. It boasts a multifaceted approach aimed at addressing Autism Spectrum Disorder (ASD) by introducing interventions and therapies in the early stages of development. This proactive stance holds immense value as it enables the reliable identification of ASD at an early stage, which in turn paves the way for enhanced developmental outcomes in individuals affected by the disorder. By harnessing the power of machine learning, the project has succeeded in devising a highly sophisticated model that can achieve the following objectives:

- 1. Individualized Treatment Programs:** The machine learning model developed in the project assists in creating individualized treatment programs based on specific traits

and needs of individuals with ASD. By adjusting interventions to the specific needs of people with ASD, this personalised approach maximises the support given to them.

2. Early Intervention: The project enables early intervention by reliably identifying ASD at an early stage. Early identification is crucial for initiating appropriate therapies and interventions, which can significantly enhance the developmental outcomes of individuals with ASD.

3. Resource Optimization: By streamlining the diagnostic process and accurately identifying individuals who require specialized assistance and services, the project ensures effective use of resources. This helps allocate resources to those who genuinely need them, optimizing support and care for individuals with ASD.

4. Screening Tool: The project provides a screening tool for ASD, which can help identify potential cases that may have been missed otherwise. By efficiently screening individuals, it expedites the diagnostic procedure and reduces the burden on medical practitioners, allowing for earlier diagnosis and intervention.

2 LITERATURE SURVEY

2.1 Existing problem

Many studies have focused on identifying the most informative features for predicting ASD. Feature selection methods, such as genetic algorithms, recursive feature elimination, and correlation-based feature selection, have been used to extract relevant features from multiple data sources. including behaviour, neuroimaging and genetics.

Various ML algorithms have been used to predict ASD, including decision trees, support vector machines (SVMs), random forests, and neural networks. These algorithms exploit the extracted features to classify individuals as ASD or non-ASD based on pattern recognition.

The studies used multiple data sources, such as behavioral assessments, questionnaires, eye tracking data, electroencephalograms (EEGs), functional magnetic resonance imaging (fMRI), and genetic markers transmission. The integration of multiple data methods has promised to improve forecasting accuracy.

Cross-validation techniques, such as k-fold cross-validation, have been used to evaluate the generalization performance of ML models. Evaluation measures such as accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC-ROC) were used to measure the predictive performance of the models.

In response to the need for interpretability in ASD predictive models, several studies have explored techniques such as feature significance analysis, model visualization, and rule

mining to provide information about the factors that contribute to the prediction results, improving the transparency and reliability of the model.

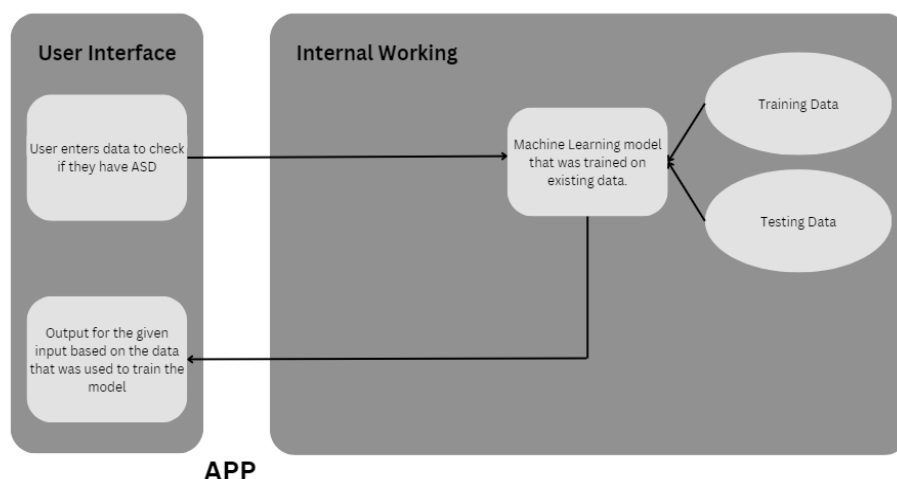
Although ML methods for ASD prediction have shown promising results, challenges remain. Limited sample sizes, data heterogeneity, and the need for standardization pose challenges in building robust and generalizable models. In addition, issues related to equipment redundancy, feature redundancy, and model complexity need special attention.

2.2 Proposed solution

- **Data Preparation:** We use a dataset that includes relevant features and corresponding labels indicating the presence or absence of ASD. We ensure the dataset is properly labeled and contains a suitable number of instances for training and evaluation. We then preprocess the data by handling missing values, removing outliers, and normalizing numerical features.
- **Feature Selection and Engineering:** We analyze the dataset and select the most informative features that are likely to contribute to ASD prediction.
- **Splitting the Dataset:** We divide the dataset into training and testing sets. Typically, around 70-80% of the data is used for training the logistic regression model, while the remaining 20-30% is reserved for evaluating its performance.
- **Model Training:** We apply logistic regression to the training data. This process involves iteratively adjusting the weights to minimize the logistic loss function, which quantified the discrepancy between predicted probabilities and actual labels.
- **Model Evaluation:** Once the logistic regression model is trained, we evaluate its performance on the testing set.
- **Prediction and Deployment:** Once the model demonstrates satisfactory performance, it can be deployed for predicting ASD in new, unseen cases. We apply the trained logistic regression model to new data instances, extract the features, and calculate the probability of ASD using the model's coefficients.

3 THEORITICAL ANALYSIS

3.1 Block diagram



3.2 Hardware / Software designing

HARDWARE

Computer: A sufficiently powerful computer with enough memory and processing capability to handle the data processing and machine learning model training. A desktop computer or a powerful laptop can be used in this.

Storage: Enough capacity to store the dataset, the results of feature extraction, trained models, and any other files or resources that may be required.

GPUs (Graphics Processing Units): Having a GPU will considerably speed up the process for more computationally demanding activities, such as training machine learning models.

SOFTWARE

Programming Language: A programming language suitable for implementing the machine learning algorithms and data processing tasks like Python.

Machine Learning Libraries: machine learning libraries and frameworks, such as scikit-learn, TensorFlow, PyTorch, or Keras, to implement and train machine learning models efficiently.

Data Processing Tools: We need tools for preprocessing, cleaning, and transforming the data. Libraries like Pandas or NumPy in Python are helpful in handling and manipulating data.

Integrated Development Environment (IDE): An IDE or code editor that supports python and offers features like syntax highlighting, debugging capabilities, and code management. Like PyCharm, Spyder, Jupyter Notebook, or Visual Studio Code.

Visualization Libraries: We also require visualization libraries, such as Matplotlib or Seaborn, to create visualizations and plots that aid in data exploration and model evaluation.

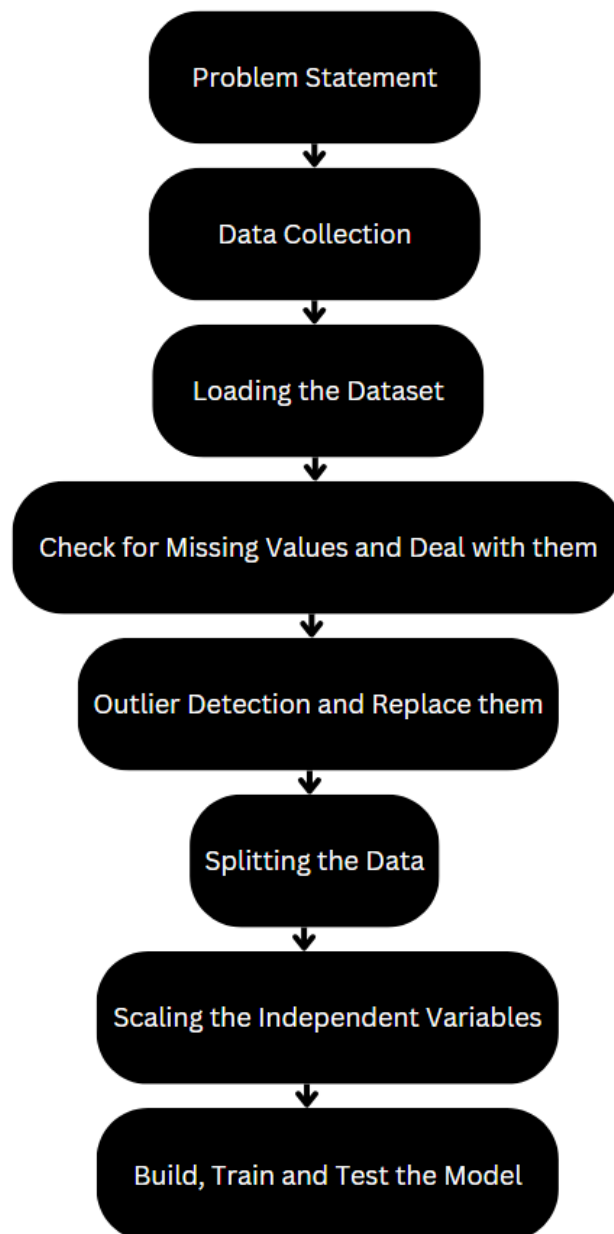
4 EXPERIMENTAL INVESTIGATIONS

- **Exploratory Data Analysis (EDA):** We perform an in-depth exploration of the dataset to gain insights into the distribution of features, identify potential outliers or missing values, and understand any patterns or relationships within the data.
- **Feature Importance Analysis:** We assess the importance of individual features in the logistic regression model. This analysis helps identify the most influential variables that contribute significantly to predicting ASD.
- **Model Interpretation:** Analyze the coefficients obtained from the logistic regression model to interpret the relationship between features and the likelihood of ASD. Positive or negative coefficient values indicate the direction and strength of the association.
- **Cross-validation and Generalization:** We conduct cross-validation to assess the model's generalizability and stability. We split the data into multiple folds and train/evaluate the model on different combinations of folds. This analysis helps verify if the model's performance is consistent across different subsets of the data and provides insights into its ability to generalize to unseen cases.

- **Comparison with Other Models:** We compare the performance of the logistic regression model with other machine learning algorithms commonly used for ASD prediction, such as decision trees, support vector machines, random forests, or neural networks. This analysis allows for an evaluation of the relative strengths and weaknesses of different models in predicting ASD.

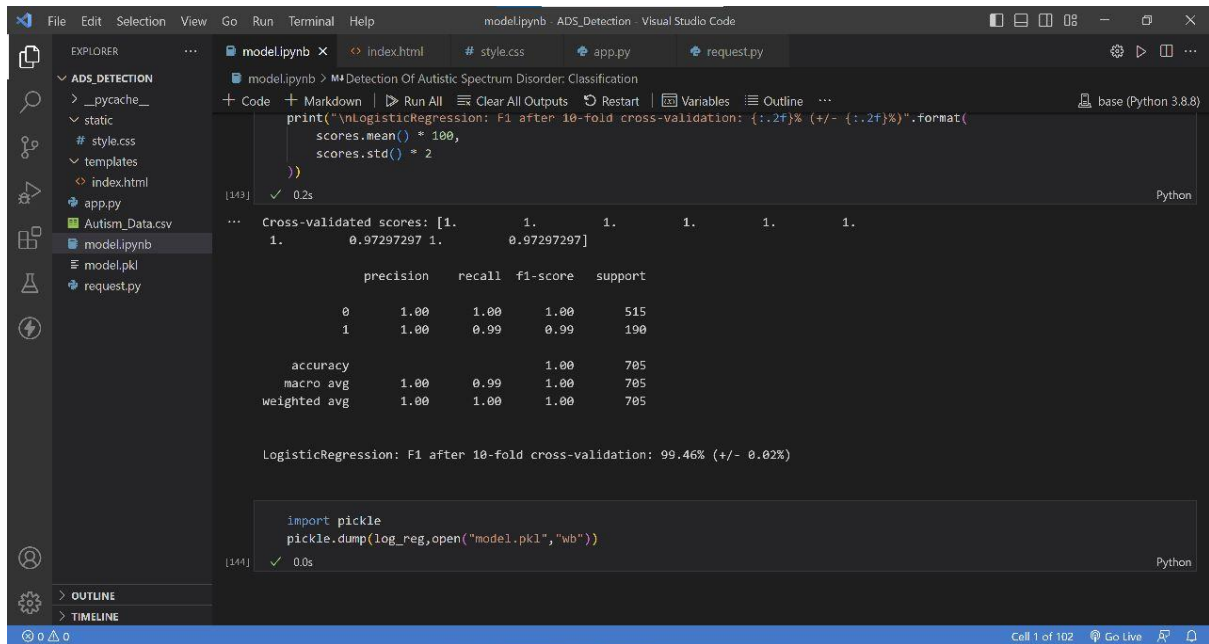
5 FLOWCHART

Diagram showing the control flow of the solution



6 RESULTS

Creation of pickle file:



The screenshot shows the Visual Studio Code interface with a file explorer on the left and a code editor on the right. The file explorer shows a project named 'ADS_DETECTION' with files like 'Autism_Data.csv', 'model.ipynb', 'model.pkl', and 'request.py'. The code editor shows the 'model.ipynb' file with a cell containing the following code:

```
print("\nLogisticRegression: F1 after 10-fold cross-validation: {:.2f}% (+/- {:.2f}%)".format(
    scores.mean() * 100,
    scores.std() * 2
))
```

The output of the cell shows the cross-validated scores and a table of metrics:

```
Cross-validated scores: [1. 1. 1. 1. 1. 1.
1. 0.97297297 1. 0.97297297]
```

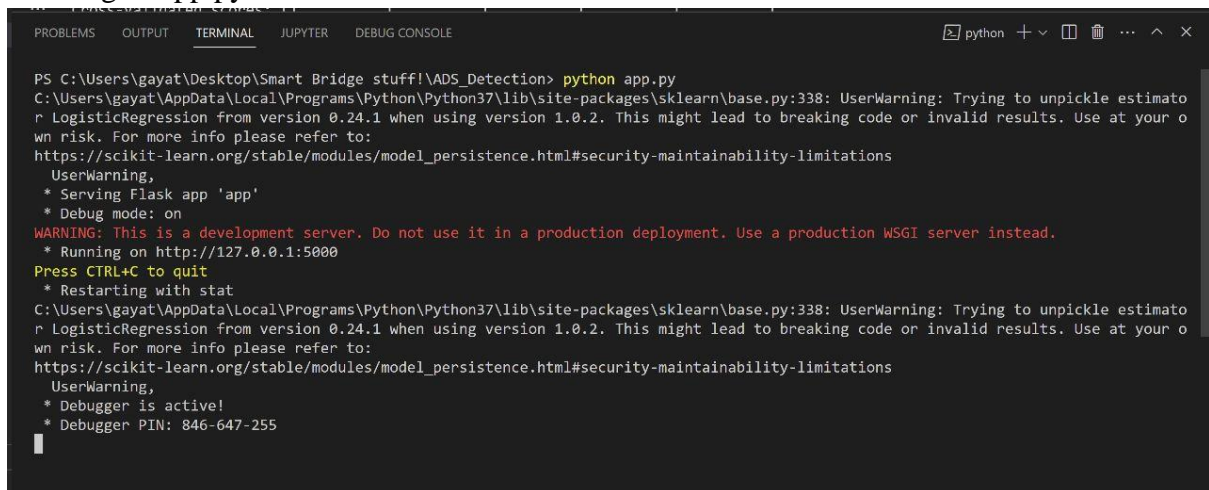
	precision	recall	f1-score	support
0	1.00	1.00	1.00	515
1	1.00	0.99	0.99	190
accuracy			1.00	705
macro avg	1.00	0.99	1.00	705
weighted avg	1.00	1.00	1.00	705

The output also shows the F1 score after 10-fold cross-validation: 99.46% (+/- 0.02%).

The code editor also shows a cell with the following code:

```
import pickle
pickle.dump(log_reg,open("model.pkl","wb"))
```

Running of app.py:



The screenshot shows a terminal window with the following output:

```
PS C:\Users\gayat\Desktop\Smart Bridge stuff\ADS_Detection> python app.py
C:\Users\gayat\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\base.py:338: UserWarning: Trying to unpickle estimator LogisticRegression from version 0.24.1 when using version 1.0.2. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
UserWarning,
* Serving Flask app 'app'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit
* Restarting with stat
C:\Users\gayat\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\base.py:338: UserWarning: Trying to unpickle estimator LogisticRegression from version 0.24.1 when using version 1.0.2. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations
UserWarning,
* Debugger is active!
* Debugger PIN: 846-647-255
```

The website:

The screenshot displays a web application titled "Autism Detection" running on a local server at 127.0.0.1:5000. The form is centered on a dark gray background and consists of a light gray container with the title and input fields. The input fields are arranged vertically and include: A1_Score, A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A7_Score, A8_Score, A9_Score, A10_Score, Age, Gender, Ethnicity, Jaundice, Country of residence, Used app before or no, Result, Age_desc, and Relation. A "Predict Autism" button is located at the bottom of the form. The browser's address bar and taskbar are visible at the top and bottom of the image, respectively.

7 ADVANTAGES & DISADVANTAGES

Advantages:

- Results from logistic regression are easily understood since the coefficients can be connected to the influence of each feature on the prediction. This makes it possible for researchers and therapists to comprehend and articulate the variables influencing ASD prediction.
- Logistic regression is a straightforward technique that is simple to comprehend, making it available to academics and practitioners of all levels of experience. Compared to sophisticated models, it has fewer hyperparameters that need to be adjusted, which makes the model creation process simpler.
- Compared to more complicated algorithms like deep learning models, logistic regression has a low computational complexity. It is excellent for applications with

limited resources since it can be trained well on huge datasets without requiring a lot of processing resources.

- Logistic regression can handle scenarios when the available dataset is modestly sized. **Robustness with Small Sample Sizes.** It can nevertheless produce somewhat accurate predictions and prevent overfitting, making it useful in situations when there is a dearth of data, such as in the case of uncommon disorders like ASD.

Disadvantages

- Logistic regression makes the assumption that there is a linear connection between the characteristics and the outcome's log-odds. Logistic regression may not successfully capture these patterns if there are complicated interactions between features or if the connection is very nonlinear, which might result in less accurate prediction.
- Logistic regression may have trouble capturing nuanced linkages and complex interactions between variables. **Limited Representation of Complex Patterns.** More sophisticated algorithms like random forests or neural networks may be able to make better predictions if the data contains high-dimensional or nonlinear correlations.
- Since logistic regression aims to minimise the logistic loss function, it may be sensitive to outliers in the dataset. Outliers that considerably differ from the majority of the data might distort the coefficient estimates of the model and perhaps have an impact on how well it performs.
- If the discriminating variables are very nuanced, logistic regression may have trouble capturing the fine distinctions between ASD and non-ASD patients.

8 APPLICATIONS

The project described, focusing on developing a machine learning model for the detection of Autism Spectrum Disorder (ASD), has several potential applications. Here are a few applications:

1. **Research and Insights:** The project's findings and the machine learning model itself can contribute to ongoing research in the field of ASD. The model's analysis of the dataset and identification of crucial factors for ASD detection can provide valuable insights into the underlying patterns and characteristics of the disorder. These insights can help researchers better understand ASD and inform the development of future diagnostic tools and interventions.
2. **Public Health Planning:** The project's machine learning model, when deployed at scale, can contribute to public health planning related to ASD. By accurately estimating the prevalence of ASD within a population, the model can assist in resource allocation, policy development, and the planning of healthcare services to meet the specific needs of individuals with ASD.
3. **Remote Screening and Telehealth:** The project's machine learning model can be integrated into remote screening and telehealth platforms. This allows for the assessment of ASD in individuals who may not have easy access to in-person healthcare services, such as those in rural or underserved areas. By leveraging the

model's predictive capabilities, remote screening and telehealth initiatives can extend the reach of ASD detection and intervention to a wider population.

4. **Public Awareness and Advocacy:** The project's findings and the utilization of the machine learning model can contribute to public awareness and advocacy efforts for individuals with ASD. By highlighting the importance of early intervention, personalized support, and optimized resource allocation, the project can help raise awareness about ASD and advocate for policies and initiatives that promote inclusivity, support, and equal opportunities for individuals with ASD.
5. **Support for Medical Practitioners:** The machine learning model can serve as a valuable tool for medical practitioners and diagnosticians. By providing accurate predictions and insights, the model can assist healthcare professionals in the diagnostic process, reducing the likelihood of misdiagnosis or missed cases. This support can enhance the confidence and decision-making of medical practitioners, leading to more effective and efficient assessments of ASD.

9 CONCLUSION

This project utilizes logistic regression to predict Autism Spectrum Disorder (ASD) based on given features. The project highlights the ease of implementation and interpretability of logistic regression, as well as its computational efficiency and potential for reasonable predictive accuracy. However, it also acknowledges the limitations of logistic regression, such as its limited model complexity, sensitivity to outliers, assumptions of independence, and inability to capture complex nonlinear relationships. considering alternative models and addressing potential limitations for more robust and accurate predictions.

After careful consideration of the characteristics of the data, the complexity of the relationships being modelled, the specific performance metrics and objectives of the project we selected logistic regression to be the best fit for this dataset. This is confirmed by evaluating its performance using appropriate metrics such as accuracy(99.46%), precision, recall, F1 score and support.

10 FUTURE SCOPE

The project on developing a machine learning model for the detection of Autism Spectrum Disorder (ASD) holds promising future scope. Here are some potential areas for future development and expansion of this project:

1. **Refinement and Enhancement of the Model:** As more data becomes available and the model is deployed in real-world settings, continuous refinement and enhancement of the machine learning model can be pursued. This includes incorporating new features, improving the accuracy and reliability of predictions, and optimizing the model's performance through iterative updates.
2. **Expansion to Other Neurodevelopmental Disorders:** The machine learning model developed for ASD detection can be extended to include the detection and

identification of other neurodevelopmental disorders, such as attention-deficit/hyperactivity disorder (ADHD), intellectual disabilities, and specific learning disorders. This expansion would provide a broader framework for early intervention and support across multiple conditions.

3. **Integration with Wearable Devices and Sensors:** The project can explore the integration of the machine learning model with wearable devices and sensors to gather real-time data on various physiological and behavioral indicators associated with ASD. This can enable a more comprehensive and dynamic approach to ASD detection and monitoring, providing valuable insights for personalized interventions and progress tracking.
4. **Development of Mobile Applications:** The project can explore the development of mobile applications that incorporate the machine learning model. These applications can serve as user-friendly interfaces, allowing caregivers, parents, and even individuals themselves to access the screening tool, personalized treatment recommendations, and progress tracking features, promoting greater engagement and involvement in the management of ASD.
5. **Collaboration with Healthcare Systems and Institutions:** Collaborations with healthcare systems, institutions, and research organizations can facilitate the integration of the developed model into existing diagnostic protocols and clinical practices. This would enhance the model's practical application and enable its widespread use in routine screening and diagnostic procedures.