Predicting Autism Spectrum Disorder Using Machine Learning Classifiers

PROBLEM STATEMENT

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects communication, social interactions, and behavior. The current diagnostic process for ASD is lengthy, often involving multiple consultations with specialists and taking several months to complete. This delay in diagnosis hampers the timely initiation of early interventions, which are crucial for improving outcomes in individuals with ASD. Traditional diagnostic methods lack the speed and efficiency required for quick detection, especially in adolescence and adulthood where the complexity of detection increases.

MATERIAL AND METHODOLOGY

Data Collection and Understanding

For this project, we utilized an open-source dataset from the UCI Machine Learning Repository, curated by Tabtah. The dataset combines demographic data and responses to the Autism Quotient (AQ) test, a 10-question assessment commonly used to evaluate autistic-like behaviors. This comprehensive dataset includes features such as attention switching, attention to details, communication, and imagination, offering valuable insights for predicting Autism Spectrum Disorder (ASD). The AQ test, developed by the Autism Research Center at Cambridge University, serves as a primary tool for identifying behavioral tendencies associated with ASD.

Features of the Dataset:

- **A1_Score:** Sensitivity to small sounds (Binary: 0,1)
- **A2_Score:** Ability to focus on the bigger picture (Binary: 0,1)
- **A3_Score:** Ease of multitasking (Binary: 0,1)
- **A4_Score:** Ability to resume tasks after interruptions (Binary: 0,1)
- **A5_Score:** Understanding implicit communication (Binary: 0,1)
- **A6_Score:** Ability to gauge if someone is bored while listening (Binary: 0,1)
- **A7_Score:** Difficulty understanding characters' intentions in stories (Binary: 0,1)
- **A8_Score:** Interest in collecting information about categories (Binary: 0,1)
- **A9_Score:** Ability to read people's emotions through facial expressions (Binary: 0,1)

- **A10_Score:** Difficulty understanding others' intentions (Binary: 0,1)
- **Age:** Age of the patient (String)
- **Gender:** Gender of the patient (Boolean: 'f' = female, 'm' = male)
- **Ethnicity:** Ethnicity of the patient (String)
- **Jaundice:** History of jaundice at birth (Boolean: 'YES', 'NO')
- **Autism:** Immediate family member diagnosed with autism (Boolean: 'YES', 'NO')
- **Country of residence:** Patient's country of residence (String)
- **Relation:** The relationship of the person who completed the test (String)
- **Class/ASD:** The target variable, indicating if the person is classified with ASD (1 = Yes, 0 = No)

Pre-Processing:

To ensure a cleaner and more focused dataset, we dropped irrelevant features such as the country of residence, age description, app usage history, and result. These features were deemed to have minimal impact on the prediction model and could introduce noise into the analysis. The remaining features were pre-processed and used to develop the machine learning models.

Approaches

In this section, we outline the machine learning models used to predict Autism Spectrum Disorder (ASD), focusing on their methodologies and significance in refining early diagnosis.

Using Logistic Regression (LR)

Logistic Regression is a widely used statistical model designed for binary classification tasks, where the outcomes are 0 or 1. LR uses a sigmoid function to model the relationship between a binary dependent variable and input features. The output is a probability score that predicts the likelihood of the binary event (ASD or not).

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(h_\theta(x_i)) + (1-y_i)\log(1-h_\theta(x_i))]$$

where hθ(is the predicted probability and yI is the actual label.

Logistic Regression is a suitable choice for binary classification problems like predicting ASD, as it provides interpretable results and is computationally efficient.

Using Random Forest (RF)

Random Forest is an **ensemble learning algorithm** that creates multiple decision trees from random subsets of data and features. The final prediction is determined by combining the results of these individual trees through majority voting (for classification) or averaging (for regression). Random Forest is particularly useful for handling complex relationships in the data and is resilient to overfitting and noise.

Gini Impurity: This measures how often a randomly selected element would be incorrectly labeled if it were randomly assigned a label from the distribution of labels in the dataset. It is used to split nodes in decision trees.

$$\text{Information Gain} = \text{Entropy(parent)} - \sum \frac{|child|}{|parent|} \times \text{Entropy(child)}$$

Information Gain: It is the reduction in entropy (or impurity) achieved by splitting a dataset based on a feature. It helps in selecting the best feature for node splits in decision trees.

Entropy : It measures the disorder or impurity in a dataset. A pure dataset has low entropy, while a highly mixed dataset has high entropy.

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

Random Forest's ability to manage high-dimensional data and its robustness make it an excellent choice for ASD prediction, where patterns are often subtle and complex.

Conclusion

In this study, we explored the potential of machine learning models to predict Autism Spectrum Disorder (ASD) using a dataset that incorporates demographic and behavioral features. By applying various machine learning algorithms such as Logistic Regression and Random Forest, we aimed to address the challenges of early ASD diagnosis.

Our models demonstrated the effectiveness of machine learning in distinguishing between individuals with and without ASD based on the Autism Quotient (AQ) test responses and other relevant features. Logistic Regression provided a straightforward and interpretable

approach, while Random Forest offered robust performance by handling complex patterns and reducing the risk of overfitting.

TEAM MEMBERS

SIDDARTH VINNAKOTA - RA2211004010396

KARTHIKEYA MADE - RA2211004010400