Autism Prediction using Machine Learning

Project Overview

The *Autism Prediction* project is a machine learning-based diagnostic support system developed to identify early signs of **Autism Spectrum Disorder (ASD)** in individuals. It leverages a clinical dataset containing demographic information and responses to standardized screening questions. The project aims to aid healthcare professionals and caregivers by providing a fast, preliminary assessment tool that can signal whether further medical consultation is necessary.

Goals and Objectives

- To develop a classification model that can predict the presence of autism traits based on input attributes.
- To compare the performance of multiple ML algorithms.
- To preprocess real-world health data for robust and generalizable model training.
- To improve diagnosis accessibility, especially in regions with limited healthcare resources.

Technologies and Tools Used

- **Python Libraries**: NumPy, Pandas, Seaborn, Matplotlib
- **Machine Learning Models**:
    - Support Vector Machine (SVM)
    - Logistic Regression
    - XGBoost Classifier
- **Preprocessing Techniques**:
    - Label Encoding
    - Feature Scaling
    - Handling Imbalanced Data with Random OverSampler
- **Evaluation Metrics**: Accuracy, Confusion Matrix, Classification Report

Dataset and Features

The dataset includes:

- **Demographic features**: age, gender, ethnicity, country of residence
- **Screening answers**: Responses to 10 behavioral screening questions
- **Other**: Family history of ASD, relation to patient, used screening app before, etc.

Target column: Class (Binary - Yes for ASD, No for non-ASD)

Workflow Steps

1. **Data Exploration & Cleaning**:
    - Loaded the dataset and identified missing/null values.
    - Cleaned and normalized categorical data using LabelEncoder.
2. **Feature Scaling & Preprocessing**:
    - Scaled continuous variables using StandardScaler to ensure consistency in training.
    - Used RandomOverSampler to handle class imbalance.
3. **Model Building**:
    - Built and trained three classifiers:
        - **Logistic Regression** – for interpretability and baseline performance.
        - **Support Vector Machine (SVM)** – for margin-based separation of classes.
        - **XGBoost** – an optimized gradient boosting framework known for high accuracy.
4. **Performance Evaluation**:
    - Used accuracy score and classification report to evaluate each model.
    - Visualized confusion matrix to identify true vs. false predictions.
    - Comparing all models to select the best performer.

Results & Insights

- **XGBoost** delivered the highest classification accuracy and balanced precision-recall performance, making it suitable for deployment in medical pre-screening tools.
- Preprocessing steps like label encoding and oversampling significantly improved model effectiveness.
- The model demonstrated that even with limited features, ML could identify early indicators of autism with acceptable accuracy.

Impact and Application

- **Potential Use**: Can be integrated into mobile apps or web-based forms for parents, schools, and clinics.
- **Scalability**: Dataset-agnostic architecture means it can be expanded with more data or integrated into broader healthcare systems.
- **Ethical Use**: Designed as an aid, not a substitute, for clinical diagnosis. Emphasizes user privacy and medical follow-up.