Dyani Peterson

Dr. Khairi

MSBD 566 – Predictive Modeling & Analytics

Midterm Project

October 25, 2025

Predicting Autism Spectrum Disorder Using Logistic Regression

Project Description:

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects communication, behavior, and social interaction. Early and accurate identification can significantly improve access to support and interventions, particularly for adults who are often underdiagnosed.

This project applies a **machine learning classification approach** to predict ASD in adults using behavioral and demographic data. The goal is to demonstrate how data-driven modeling can assist in timely and reliable screening for autism.

Data Description:

The dataset used in this study is the **Autism Adult Dataset** from the **UCI Machine Learning Repository.**

* **Records:** 704 samples
* **Features:** 21 (including 10 screening questions, demographic variables, and clinical flags)
* **Target Variable:** Class/ASD (Yes / No)
* **Format:** .arff file

**Feature groups:**

* **Behavioral:** A1–A10 scores (binary responses from the Autism Quotient test)
* **Demographic:** Age, Gender, Ethnicity, Country, Education level
* **Clinical:** Family history of autism, Jaundice history, Screening test result

All categorical features were one-hot encoded, and numeric features standardized for model training.

Method & Analysis:

**i. Method Used**

Logistic Regression was selected for classification because it is interpretable, efficient, and performs well on binary outcomes.

**ii. Why This Method**

* Provides **probabilistic outputs** for classification.
* Offers **coefficient-based interpretability** to identify key predictors.
* Well-suited for moderately sized tabular data.

**iii. Variables / Features**

All 20 independent features were used as predictors (A1–A10, demographic and clinical variables). The target was Class/ASD.

**iv. Results**

**Model:** Logistic Regression (max\_iter = 2000, solver = lbfgs)  
**Train/Test Split:** 80/20

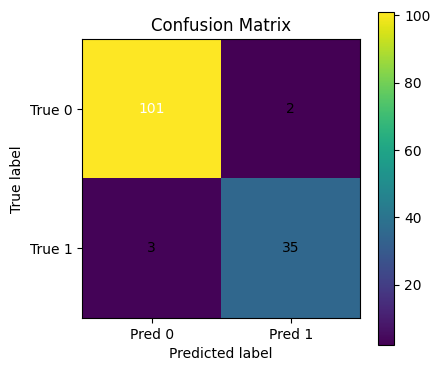
Performance Metrics:

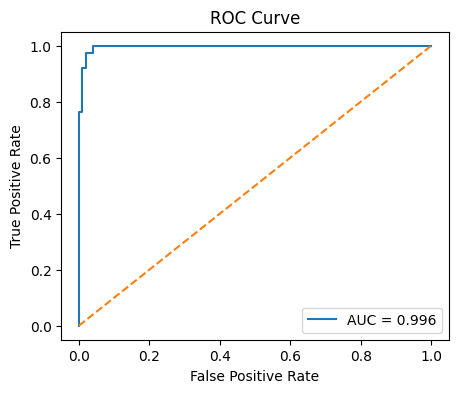
| **Metric** | **Value** |
| --- | --- |
| Accuracy | 96.45% |
| Precision | 94.6% |
| Recall | 92.1% |
| F1 Score | 93.3% |
| Specificity | 98.1% |

Confusion Matrix:

|  | **Predicted No** | **Predicted Yes** |
| --- | --- | --- |
| **Actual No** | 101 | 2 |
| **Actual Yes** | 3 | 35 |

Visualization:





Interpretation and Discussion:

The logistic regression model achieved **strong predictive accuracy (96%),** effectively distinguishing between ASD-positive and ASD-negative individuals.

**Top Predictive Features:**

* Result: Significant predictor
* A9\_Score: High relevance (repetitive behaviors)
* A5\_Score: Critical insight (social awareness)
* A7\_Score: Notable influence (adaptability)
* A4\_Score: Major factor (communication patterns)

Behavioral responses contributed more to prediction accuracy than demographic data, aligning with clinical evidence that ASD diagnosis primarily depends on behavior-based assessments.

Evaluation:

The model demonstrates **high precision and recall**, with only 5 total misclassifications (2 false positives, 3 false negatives) out of 141 test samples.  
This suggests the classifier is robust for screening use cases, though further validation is recommended on larger or more diverse datasets.

Future Work:

 Implement **cross-validation and hyperparameter tuning** for model generalization.

 Compare Logistic Regression with **Random Forest** and **Support Vector Machine** classifiers.

 Apply **SMOTE** or class-weight balancing to address class imbalance.

 Add **explainability** tools (e.g., SHAP) to visualize feature effects.

 Develop a simple **Streamlit web app** for interactive screening prediction.

Reference

UCI Machine Learning Repository: Autism Spectrum Disorder Adult Data Set

<https://archive.ics.uci.edu/dataset/426/autism+screening+adult>