**EXECUTIVE SUMMARY FOR AUTISM PREDICTION**

The dataset consists of 800 entries with features related to autism spectrum disorder (ASD) prediction, including binary questionnaire scores (A1\_Score to A10\_Score), demographic information (age, gender, ethnicity, country of residence, etc.), medical history (jaundice, autism related), application usage, and the target variable "Class/ASD" indicating ASD diagnosis.

**Data Understanding and Preprocessing**

There are 22 columns initially; key relevant features include behavioral scores, age, gender, ethnicity, jaundice presence, autism history, and country of residence.

Data has no missing values or duplicates.

Age was converted to an integer type, and some categorical inconsistencies in ethnicity and country names were standardized.

The original dataset was filtered to keep entries from major countries for better model focus, narrowing the dataset to 635 entries.

**Exploratory Data Analysis (EDA)**

Age distribution is right-skewed (Skewness ~1.14) with some outliers—39 outliers detected and replaced with median values.

Autism cases constitute about 20% of the dataset.

Categorical features such as jaundice, autism history, gender, and ethnicity were visualized through count and stacked bar plots.

Relationships between gender and ethnicity were examined to understand demographic patterns.

**Data Preparation for Modeling**

Target variable is imbalanced (639 non-ASD vs 161 ASD).

Categorical variables were encoded using One-Hot Encoding, and numerical variables (age, result) were scaled using StandardScaler.

The processed dataset analyzed had 40 features post-encoding and scaling.

Data was split into training (80%) and test (20%) sets.

**Model Training and Evaluation**

Four ML models were trained with class weights to address imbalance: Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost.

Logistic Regression achieved ~86.6% accuracy with recall for the ASD class around 77%.

SVM showed 83.5% accuracy but lower ASD recall (~57%), indicating more false negatives.

Random Forest matched Logistic Regression at ~86.6% accuracy with better ASD precision (81%) but variable recall (57%).

XGBoost had 83.5% accuracy but lower recall (53%) on ASD class.

**Hyperparameter Tuning using Optuna**

Optuna was applied for automated hyperparameter tuning focusing on maximizing recall.

The best performing model was Random Forest with hyperparameters:

**n\_estimators=99, max\_depth=3, min\_samples\_split=5, min\_samples\_leaf=10, bootstrap=True, class\_weight={0:15,1:85}**

This tuned Random Forest model improved recall to ~92.9%.

Evaluation on train and test sets showed balanced performance with high recall on the ASD class (93% on test), confirming the model’s effectiveness for detecting ASD.

**Summary**

This project systematically explored an ASD dataset to build predictive models. The data was preprocessed and analyzed thoroughly, followed by training multiple classifiers with attention to imbalanced classes. Automated hyperparameter tuning identified an optimized Random Forest model that effectively balances recall and precision, prioritizing the detection of ASD cases accurately. This model shows promise for assisting early ASD screening based on questionnaire data and demographic/medical history features.