**Project Overview**

This project leverages machine learning to predict Autism Spectrum Disorder (ASD) in individuals based on key features from behavioral, demographic, and historical screening data. The goal is to enhance the early detection of ASD and support healthcare professionals in diagnosis.

**Objectives**

* Develop a predictive model for identifying ASD.
* Address class imbalances in the dataset using synthetic techniques.
* Analyze feature importance to better understand key indicators of ASD.

**Dataset Overview**

The dataset includes 800 samples and 22 columns, with variables such as:

* **Behavioral Scores:** A1\_Score through A10\_Score (binary scores based on specific criteria).
* **Demographics:** Gender, Age, Ethnicity, Country of Residence.
* **Medical History:** Jaundice, Autism in Family (Yes/No), Screening Test Results.
* **Target Variable:** Class/ASD (0 for no ASD, 1 for ASD).

**Data Preprocessing**

* **Handling Missing Data:** Replaced missing values in categorical fields like ethnicity and relation with meaningful defaults such as "Others."
* **Standardization:** Country names were normalized for consistency.
* **Label Encoding:** Categorical features were encoded for compatibility with machine learning models.
* **Outlier Treatment:** Applied Interquartile Range (IQR) method to handle outliers in age and result columns.

**Exploratory Data Analysis (EDA)**

* **Distribution Analysis:** Visualized distributions of age and screening results using histograms, noting skewness and outlier patterns.
* **Correlation Heatmap:** Highlighted key relationships, such as a strong positive correlation between result and Class/ASD.
* **Class Imbalance:** Addressed imbalance with 639 (Class 0) and 161 (Class 1) using SMOTE.

**Model Development**

1. **Models Implemented:**
   * Decision Tree Classifier
   * Random Forest Classifier
   * XGBoost Classifier
2. **Data Split:**
   * Training and Testing: 70%-30% split.
   * Balanced the training set using SMOTE.
3. **Cross-Validation:**
   * Performed 5-fold cross-validation for model robustness.
   * Accuracy results:
     + Decision Tree: 86%
     + Random Forest: 91%
     + XGBoost: 90%

**Key Insights and Findings**

1. **Top Predictive Features:** The result feature is the strongest predictor of ASD.
2. **Model Performance:** Random Forest provided the best performance in terms of accuracy and consistency across folds.
3. **Behavioral Scores:** Some scores (e.g., A1\_Score and A2\_Score) showed moderate positive correlations with ASD likelihood.
4. **Demographics:** Gender, ethnicity, and age did not significantly affect the prediction outcome, showing that the behavioral scores and medical history are more critical.

**Conclusion**

The Autism Prediction project demonstrates the power of machine learning in diagnosing ASD. The Random Forest model proved to be the most effective, with a cross-validation accuracy of 91%. This framework can be further enhanced by incorporating additional data sources or refining the feature selection process.