**Executive Summary: Autism Prediction Using Machine Learning**

**Introduction**

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects communication, behavior, and social interactions. Early diagnosis and intervention can significantly improve outcomes for individuals with ASD. This project aims to develop a predictive model for autism using machine learning techniques applied to a dataset containing behavioral and demographic indicators.

**Dataset and Preprocessing**

The dataset used consists of 800 records with 22 features, including responses to a screening questionnaire, demographic factors, and previous autism diagnoses. Preprocessing steps involved:

* Handling missing values and inconsistencies.
* Encoding categorical variables using Label Encoding.
* Removing outliers in age and result variables using the Interquartile Range (IQR) method.
* Standardizing numerical features using StandardScaler.
* Addressing class imbalance with Synthetic Minority Oversampling Technique (SMOTE), balancing the ASD-positive and ASD-negative cases in training data.

**Exploratory Data Analysis (EDA)**

EDA provided insights into the dataset, including:

* Distribution of ASD cases across different demographics.
* Correlation analysis to identify the most relevant features.
* Visualization of variable distributions through histograms, box plots, and scatter plots.
* Identification of potential biases or gaps in the dataset.

**Model Training and Evaluation**

Several machine learning models were trained and tested:

1. **Logistic Regression:** Achieved an accuracy of 81.9%.
2. **Support Vector Machine (SVM):** Performed best with 84.4% accuracy.
3. **Decision Tree:** Lower accuracy (77.5%) due to overfitting.
4. **Random Forest:** Achieved 82.5% accuracy with improved generalization.
5. **XGBoost:** Best model with 85% accuracy, balancing precision and recall effectively.

**Model Selection and Hyperparameter Tuning**

To improve performance, hyperparameter tuning was applied using RandomizedSearchCV on all models. The best-performing model was an optimized **SVM with RBF kernel** (C=100, gamma='auto'), achieving a cross-validated accuracy of 93% on training data and 80% on test data.

**Key Findings and Conclusion**

* The behavioral screening scores and demographic factors were strong predictors of ASD likelihood.
* Data preprocessing and feature scaling significantly improved model accuracy.
* SMOTE effectively addressed class imbalance, ensuring better generalization.
* The **SVM model** was selected as the final model due to its superior performance in handling complex decision boundaries.
* The methodology can be extended for real-world applications, aiding in preliminary autism screening.