Autism Spectrum Disorder Prediction

A Machine Learning Approach

Supervised by: Usman Ali

Name	Roll Number
Badar Rasheed Butt	211370113
Moaz Aslam	211370125
Ayesha Kanwal	211370178
Shamama Tarif	211370169

February 21, 2025

Gift University - Department of Computer Science

Contents

1 Introduction

Autism Spectrum Disorder (ASD) is a developmental condition affecting communication, behavior, and social interaction. Early diagnosis is crucial, yet traditional methods are often time-consuming and require expert evaluation. This project focuses on predicting ASD by leveraging machine learning techniques. The primary objective is to develop a custom Random Forest classifier using a dataset of 800 survey responses that include behavioral, demographic, and medical history data.

This report comprehensively discusses:

- Data preprocessing techniques (handling missing values, outlier detection, feature encoding)
- Exploratory Data Analysis (EDA) (data visualization and statistical insights)
- Custom Machine Learning Model Implementation (Decision Tree and Random Forest)
- Performance Metrics (Accuracy, Precision, Recall, F1-score)
- Evaluation Results and Future Improvements

2 Dataset Overview

The dataset comprises 22 features and 800 records. The information covered includes:

- **Survey Scores:** A1_Score to A10_Score
- **Demographic Information:** age, gender, ethnicity, country_of_residence
- **Medical History:** jaundice, autism, relation, used_app_before
- **Target Variable:** Class/ASD (1 = ASD diagnosed, 0 = Not diagnosed)

2.1 Dataset Summary

- **Total Entries:** 800
- **Numerical Features:** age, result

- **Categorical Features:** gender, ethnicity, jaundice, autism, relation, used_app_before
- **Class Distribution Before Balancing:**
  - ASD Cases (1): 230
  - Non-ASD Cases (0): 570
- **Class Imbalance:** Resolved using SMOTE.

3 Data Preprocessing & Feature Engineering

3.1 Handling Missing Values

- Missing values were identified in *ethnicity* and *relation*.
- Replaced missing entries with **"Others"** for consistency.

3.2 Feature Engineering

- **Dropped Unnecessary Features:**
  - `ID` (not useful for prediction)
  - `age_desc` (contained only one unique value)
- **Fixed Data Inconsistencies:** Corrected country names (e.g., *"Viet Nam"* → *"Vietnam"*, *"Hong Kong"* → *"China"*).

3.3 Handling Outliers

- Outliers in *age* and *result* were detected using the Interquartile Range (IQR) method.
- Outliers were replaced with median values to avoid skewing predictions.

3.4 Addressing Class Imbalance

- **Before SMOTE:** ASD Cases (230), Non-ASD Cases (570)
- **After SMOTE:** ASD Cases (570), Non-ASD Cases (570) resulting in a balanced dataset.

4 Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

- Histograms for *age* and *result* indicated right-skewed distributions.
- Box plots helped in identifying outliers in these features.

4.2 Correlation Analysis

- A heatmap was used to visualize feature correlations.
- No features exhibited high inter-correlation, ensuring minimal redundancy.

4.3 Categorical Feature Analysis

- Count plots for categorical features revealed unequal distributions.
- Notably, a higher frequency of males was observed in the dataset.

5 Machine Learning Model Implementation

5.1 Custom Decision Tree Implementation

- **Splitting Criterion:** Gini Impurity.
- **Recursive Growth:** Each node splits until the stopping criteria (max_depth or min_samples_split) are met.

5.2 Custom Random Forest Implementation

- **Number of Trees:** 50.
- **Max Depth:** 10.
- **Feature Subset Ratio:** 60% of features per tree.
- **Bootstrap Sampling:** Each tree is trained on a randomly selected subset of the data.

5.3 Model Training

- **Train-Test Split:** 80% training and 20% testing.
- **Feature Scaling:** Not required as tree-based models handle raw values efficiently.

6 Model Evaluation & Results

6.1 Performance Metrics

After training on the balanced dataset, the model was evaluated on the test set. The key results are as follows:

Confusion Matrix:

	Predicted Negative (0)	**Predicted Positive (1)**
Actual Negative (0)	107	17
Actual Positive (1)	9	27

Classification Report:

- **Precision:**
  - Non-ASD (0): 92%
  - ASD (1): 61%
- **Recall:**
  - Non-ASD (0): 86%
  - ASD (1): 75%
- **F1-Score:**
  - Non-ASD (0): 89%
  - ASD (1): 67%
- **Overall Accuracy:** 83.75%

6.2 Interpretation of Results

- The high precision for non-ASD cases suggests that the model effectively avoids false positives.
- The moderate precision for ASD cases indicates some non-ASD cases were misclassified.
- The recall for ASD cases (75%) shows that the model is reasonably good at detecting true ASD cases.
- An overall balanced accuracy of 83.75% demonstrates robust generalization.

6.3 Model Saving

The trained Random Forest model was saved as `custom_rf_model.pkl` using Python's `pickle` module.

7 Insights, Challenges, and Future Improvements

7.1 Key Findings

- The initial class imbalance was effectively mitigated using SMOTE.
- Outlier detection and replacement enhanced the overall model performance.
- The custom Random Forest classifier achieved a strong accuracy of 83.75%.

7.2 Challenges Faced

1. **Class Imbalance:** The initial bias towards non-ASD cases required oversampling.
2. **Precision for ASD Cases:** Some ASD cases were misclassified, leading to lower precision.
3. **Outlier Impact:** Outliers had to be carefully handled to avoid skewing the predictions.

7.3 Suggested Improvements

- **Hyperparameter Optimization:** Fine-tune parameters such as `max_depth`, `min_samples_split` and the number of trees.
- **Feature Engineering:** Explore interaction features and consider Principal Component Analysis (PCA) for noise reduction.
- **Algorithm Exploration:** Evaluate other models such as XGBoost, Gradient Boosting, or deep learning approaches (MLP/CNN) for improved performance.
- **Deployment:** Develop a web-based tool to deploy the model for real-world screening.

8 Conclusion

This project successfully implemented a custom Random Forest classifier for the prediction of Autism Spectrum Disorder. With an accuracy of 83.75%, the model demonstrates strong potential as an early screening tool. Future improvements through hyperparameter tuning, advanced feature engineering, and exploring alternative algorithms could further enhance its predictive capability, providing valuable support to medical professionals.