ML Report

Veermata Jijabai Technological Institute(VJTI)

Subject: Autism Prediction Using Machine Learning
Date: 09/05/2025
Professor: Tabassum Kazi

Team Members:

1. Siddhi Parekh (221071047)
2. Druhi Phutane (221071052)
3. Richa Sawant (221071058)
4. Megha Wadher (221071077)

Table of Contents

    - Classification Report

9. **Model Serialization**
    - Saving the Best Model
    - Saving Encoders

10. **Next Steps**
    - Predictive System Development
    - Performance Improvements
    - Potential Deployment (e.g., Web App)

11. **Conclusion**
    - Summary of Findings
    - Final Remarks

1. Introduction

Objective

The purpose of this project is to build a machine learning model that can predict the presence of Autism Spectrum Disorder (ASD) in individuals using screening results and demographic data.

Problem Statement

ASD is a developmental disorder that affects communication and behavior. Early identification is crucial. By leveraging machine learning on survey-based data, we aim to assist healthcare professionals in preliminary screening for autism.

---

2. Dataset Overview

Data Source

- The dataset was loaded from a CSV file named `train.csv`.
- It contains responses to a ten-question screening test and demographic information.

Feature Description

- **Numerical:** `age`, `result`
- **Categorical:** `gender`, `ethnicity`, `country of residence`, `used_app_before`, `jaundice`, `austim`, `relation`
- **Screening Scores:** `A1_Score` to `A10_Score`

Target Variable

- `Class/ASD`: Binary classification — 1 for ASD-positive, 0 for negative.

---

3. Data Preprocessing

Initial Inspection

- Dataset loaded into a Pandas DataFrame.
- Shape: (number of rows × columns)
- Displayed head/tail, data types, and column structure.

Data Cleaning

- **Dropped columns:**
  - `ID`: Unique identifier, not predictive.
  - `age_desc`: Contained only one unique value.
- **Country name normalization:** Corrected country names like "Viet Nam" to "Vietnam".
- **Missing Values:**
  - Replaced `"?"` and uncommon values in `ethnicity` and `relation` with `"Others"`.

Outlier Detection & Treatment

- Outliers in `age` and `result` detected using the IQR method.
- Replaced outliers with median values to reduce model sensitivity.

Label Encoding

- Applied `LabelEncoder` to all categorical features.
- Saved encoders using `pickle` to ensure consistent transformation during inference.

4. Exploratory Data Analysis (EDA)

Univariate Analysis

- **Numerical Features:**
  Histograms and box plots showed:
  - Age was normally distributed with minor outliers.
  - Result scores had a relatively uniform spread.
- **Categorical Features:**
  Count plots showed class imbalance in features like `gender`, `ethnicity`, and `Class/ASD`.

Bivariate Analysis

- **Correlation Matrix:**
  A heatmap showed weak to moderate correlations, indicating no multicollinearity.

Key Insights

- Dataset had imbalanced target distribution.
- Categorical classes also showed imbalance.
- No redundant highly correlated features.

---

5. Train-Test Split & Class Balancing

Train-Test Split

- Used an 80/20 split for training and testing respectively.

SMOTE (Synthetic Minority Over-sampling Technique)

- Applied to training data to balance minority class.
- Resulted in equal class distribution, which helps improve model generalization.

---

6. Model Building

Selected Algorithms

Three tree-based classifiers were trained:

- **Decision Tree**
- **Random Forest**
- **XGBoost**

Cross-Validation

- Applied 5-fold cross-validation to assess performance consistency.
- Reported mean accuracy for each model.

---

7. Hyperparameter Tuning

Parameter Grids

- Defined hyperparameter grids for:
  - Decision Tree: `max_depth`, `criterion`, etc.
  - Random Forest: `n_estimators`, `max_depth`, etc.

    - XGBoost: `learning_rate`, `subsample`, `n_estimators`, etc.

RandomizedSearchCV

- Used for efficient hyperparameter tuning with 5-fold CV.
- Selected the best estimator for each model.

Best Model Selection

- Compared best cross-validated scores.
- Choose the model with the highest score as the final classifier.

---

8. Model Evaluation

Test Set Evaluation

- Used the best model to predict on the test set.
- **Metrics Reported:**
    - **Accuracy:** Measured correctness.
    - **Confusion Matrix:** Showed TP, TN, FP, FN.
    - **Classification Report:** Included precision, recall, and F1-score for each class.

---

9. Model Serialization

Saving the Model

- The best model was saved as `best_model.pkl`.

Saving the Encoders

- Label encoders used during preprocessing were saved as `encoders.pkl` for future predictions.

---

10. Next Steps

Predictive System Development

- Use the saved model and encoders to build a prediction pipeline (e.g., Flask or Streamlit app).

Performance Improvements

- Explore ensemble techniques, feature engineering, or additional datasets.

Deployment

- Web app to allow users to input screening responses and receive predictions.

---

11. Conclusion

- A robust model for autism prediction was developed using tree-based classifiers.
- SMOTE helped to balance the data and improve accuracy.
- The best-performing model was saved and is ready for deployment.
- With further improvements and real-world validation, this model can serve as a useful tool in early ASD screening.

12. Links

Dataset Link : https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults

Github link: https://github.com/richa-sawant/ML_Project_AutismDetection

Research paper links: https://www.sciencedirect.com/science/article/pii/S101836472400380X

https://www.sciencedirect.com/science/article/pii/S2772442524000819