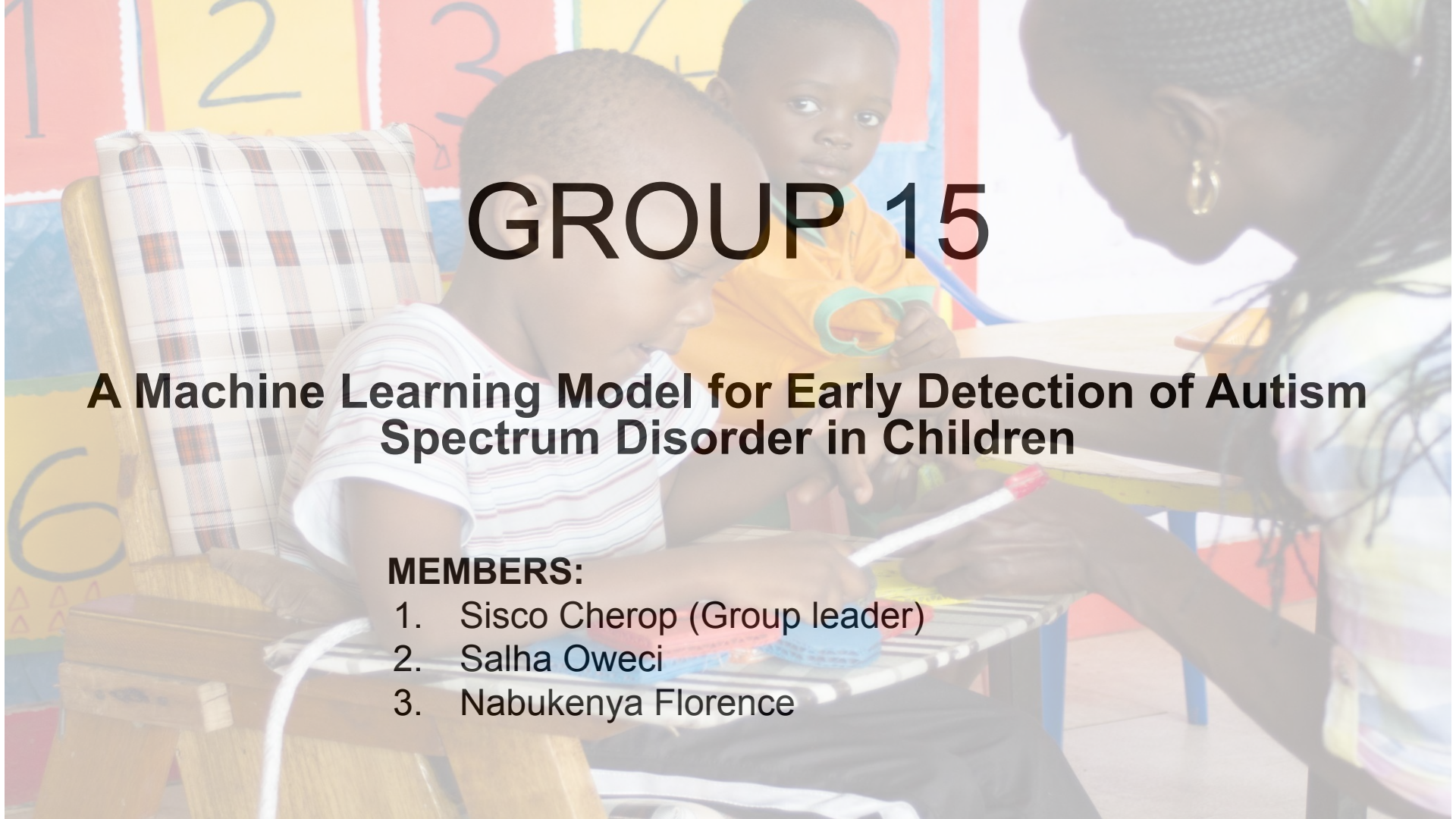


GROUP 15

A Machine Learning Model for Early Detection of Autism Spectrum Disorder in Children

MEMBERS:

1. Sisco Cherop (Group leader)
2. Salha Oweci
3. Nabukenya Florence

PROBLEM DESCRIPTION

Autism Spectrum Disorder (ASD) often goes undetected in early childhood, especially in low-resource settings like Uganda where access to specialized assessment tools and trained clinicians is limited. Many children are identified late ,sometimes at school age,after developmental delays have already affected learning, communication, and social interaction.

The major challenge is that early signs of ASD are often subtle, vary from child to child, and depend heavily on caregivers recognizing specific behavioral patterns. With rising ASD cases globally, there is an urgent need for simple, accessible, and scalable tools that can help identify children who may require further assessment.

Our project addresses this challenge by developing a machine learning based screening model that predicts ASD traits using behavioural and developmental data. This allows for earlier recognition of potential symptoms, even in environments where formal clinical evaluation is not readily available.

PROBLEM BEING SOLVED

In Uganda, the prevalence of ASD is on the rise, yet many children remain undiagnosed or misdiagnosed. A 2023 report by the Uganda Autism Society indicates that over 60% of children with ASD are diagnosed late due to limited access to specialised services and low awareness levels.

Early diagnosis and tailored support can significantly improve outcomes for children with ASD helping them thrive both academically and socially.

Our Goal

Develop an **accurate, accessible, data-driven screening tool** using behavioural, developmental, and clinical features to:

- Support early identification so that diagnosis is made earlier.
- Reduce diagnostic delay in Uganda and similar contexts.

Impact of Early Detection: Children who receive support by being diagnosed will therefore get appropriate treatment hence reducing the rise of autistic children.

DATASET OVERVIEW

Type of Dataset

- Structured tabular data which is ideal for supervised binary classification
- Each row = One child
- Each column = Behavioral, developmental, clinical, or demographic feature

Original Dataset

- 1,985 children
- 31 columns (including redundant & noisy features)
- Target Variable - ASD_traits (Yes/No) — indicates whether the child is diagnosed/likely autistic (Yes/No or 1/0)

Source of Dataset and its unique: The dataset is from Kaggle. We added two new numerical columns into our dataset based on feedback from our facilitator about adding more data since most of our initial dataset was categorical.

Github repo here

DATASET

Key categories:

- Developmental milestones
- ASD behavioural traits (A1–A10)
- Medical & family history

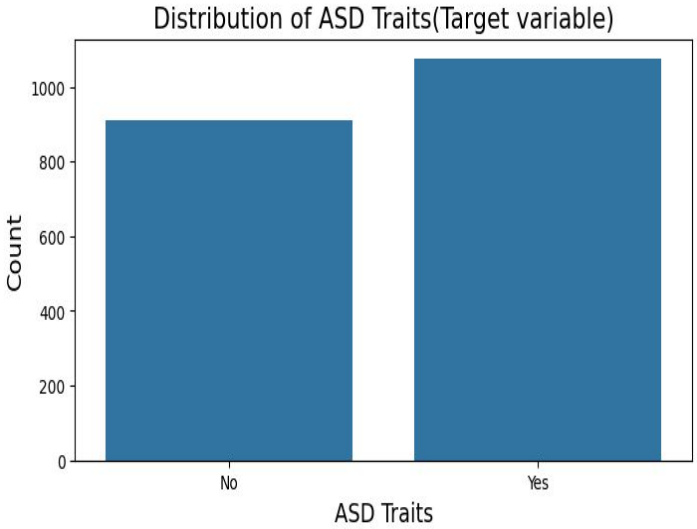
Target: ASD_traits (Yes/No or 1/0)

DATE CLEANING: After Rigorous Pre-processing & Cleaning of our dataset, we removed noisy, redundant, and unstable columns e.g., duplicate birth weight, malformed A10 ranged, Who_completed_the_test and Ethnicity. We also fixed missing values, converted all answers into numeric format, and preparing the data so the models could learn from it accurately

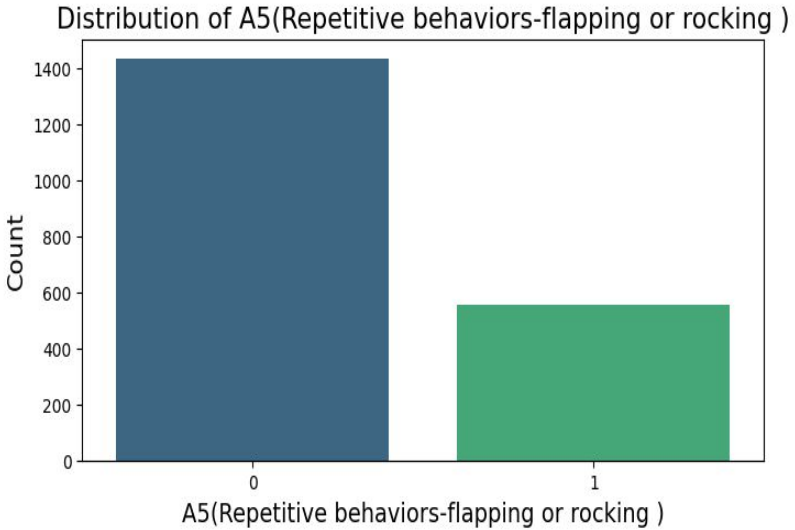
Final clean dataset: 1,950+ rows × 22 high-quality features

Why we reduced our dataset? We reduced our dataset to make our model more accurate, reliable, and easier to interpret. Some columns were noisy, duplicated, inconsistent, or not useful for predicting ASD. Keeping them would have added confusion and reduced model performance.

DATA EXPLORATION AND VISUALIZATION

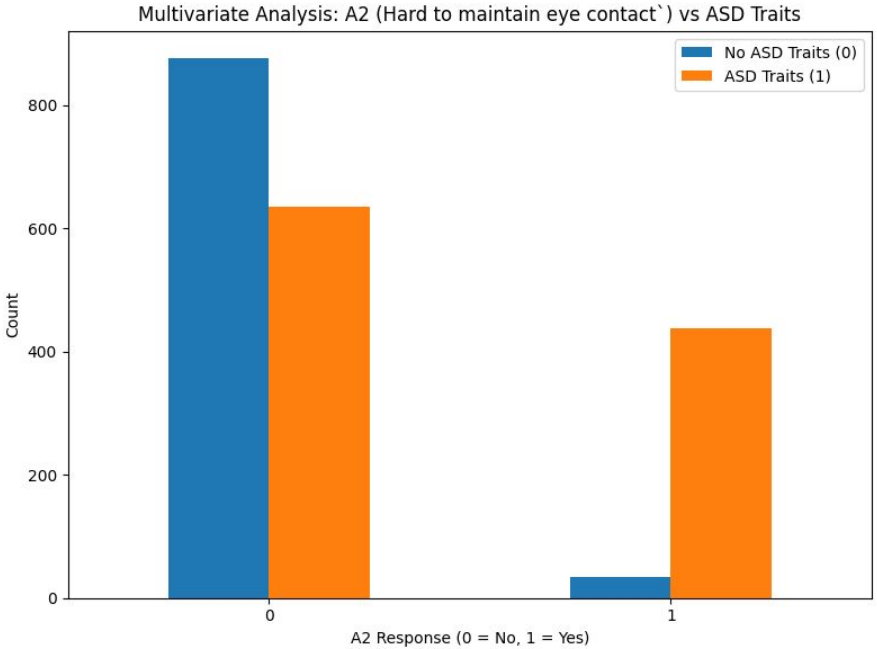


The univariate analysis shows that 54% of children screened positive for ASD traits and 46% screened negative. This is a fairly balanced distribution, which supports reliable model training. The slightly higher number of ASD-positive cases reflects the screening-focused nature of the dataset.



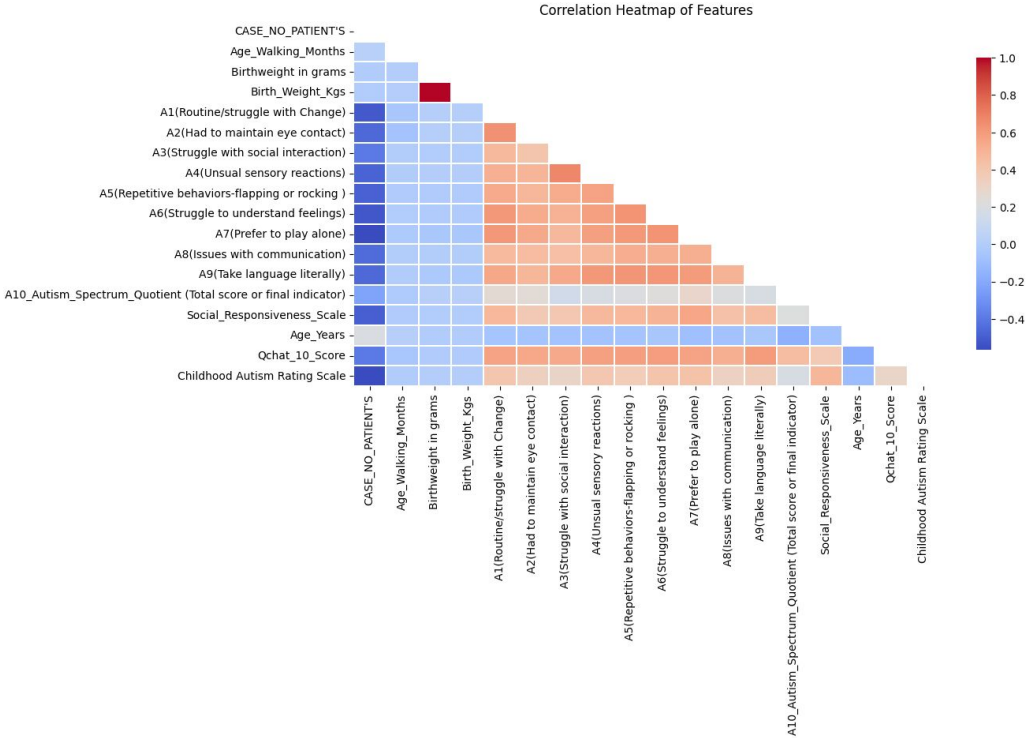
The univariate analysis of A5 shows that most children in the dataset do not display repetitive behaviors such as flapping or rocking, while a smaller portion do. This means repetitive behaviors are less common in this sample, but the number of “Yes” cases is still sufficient for the model to learn from.

DATA EXPLORATION AND VISUALIZATION



Children who struggle with eye contact are far more likely to show ASD traits, while children without eye-contact difficulties mostly fall in the non ASD group. This makes eye contact (A2) one of the strongest behavioural indicators of ASD in our dataset.

DATA EXPLORATION AND VISUALIZATION



The heatmap shows that ASD traits are strongly associated with behavioral indicators such as difficulty with eye contact, repetitive behaviors, communication issues, and taking language literally. Developmental conditions such as speech delay, learning disorders, and anxiety also show moderate correlation. Demographic factors such as age, birth weight, and ethnicity have weak or negative correlation, meaning they contribute minimally to ASD prediction. This validates that our dataset appropriately captures the key behavioral markers of ASD, making it suitable for machine learning classification

MACHINE LEARNING

We trained Two types of models:

1. **Random Forest Classifier**

We used Random Forest because it works very well with binary data like ours.

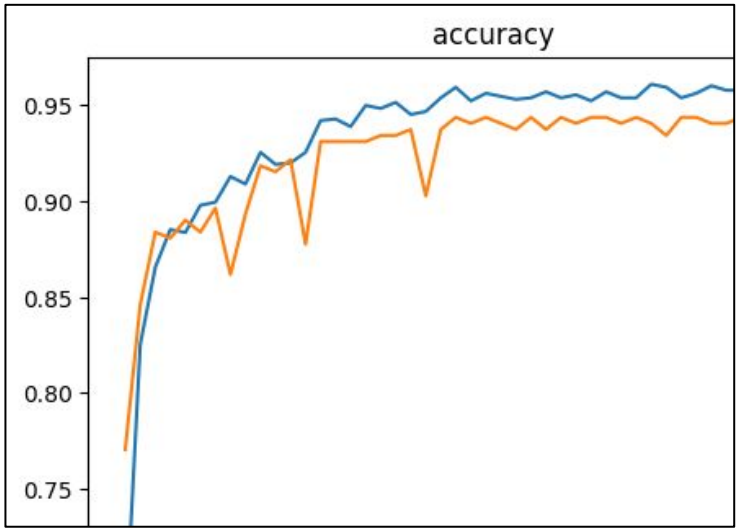
- It is good for high stability and interpretation especially for which features matter.
- Handles nonlinear patterns very well
- Handles complex behaviour patterns
- Excellent for tabular dataset
- No scaling needed

2. **Neural Network (Deep learning)**

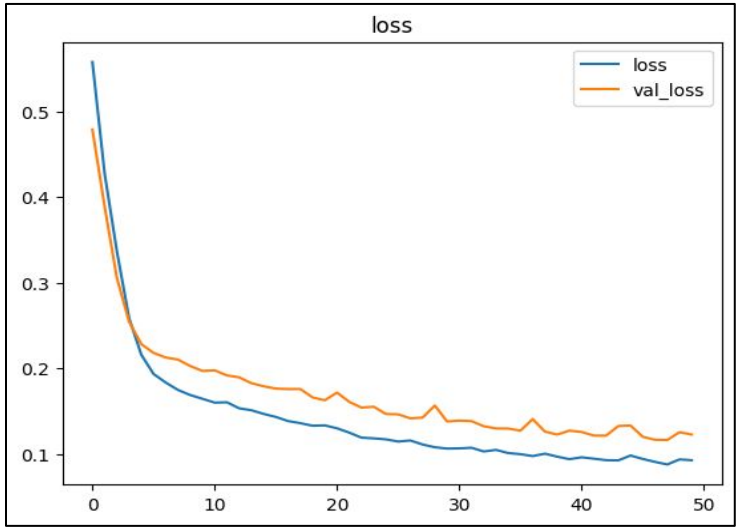
- It learns patterns the model creates on its own like a human brain. (Can theoretically learn very complex patterns)
- Worked as accurate as the Random forest which showed that our model was strong.

(Deep Learning)

INTERPRETATION OF ACCURACY AND LOSS GRAPHS



The model performs well on data it has never seen. The validation line is close to the training line meaning no overfitting and the model is accurate and stable.



The model is learning effectively without overfitting. Training and validation loss are both low with similar pattern meaning model works well with new data

MODEL RESULTS

	RANDOM FOREST	NEURAL NETWORK
Accuracy	96.22%	92.95%
Precision	0.97%	0.93%
F1-Score	0.96%	0.94%
Recall	0.96%	0.94%
Confusion Matrix	[[165, 17]\n [13, 202]]	[[168, 14]\n [13, 202]]

- Both models performed very well, but the Random Forest achieved slightly higher scores across all metrics.
- High recall means our model successfully found almost all ASD-positive children, with very few missed cases.
- Precision shows how accurate the model is when it predicts ASD.High precision means very few false alarms.
- The F1-score is a balanced measure of both precision and recall, showing how well the model identifies ASD cases without making too many mistakes.

Neural Network performs well but Random Forest remains best.

CONCLUSION

- Our ASD prediction model is accurate, reliable, and practical.
- Our model can be used to support early ASD screening in schools, public health, parents and caregivers, hospitals, communities,health cares and across uganda.
- Helps clinical experts focus early on children with high-risk of ASD.

Future work: We hope to integrate this into mobile tools, larger datasets.

MODEL DEPLOYMENT : here

MEMBERS:

1. Sisco Cherop (Group leader)
2. Salha Oweci
3. Nabukenya Florence