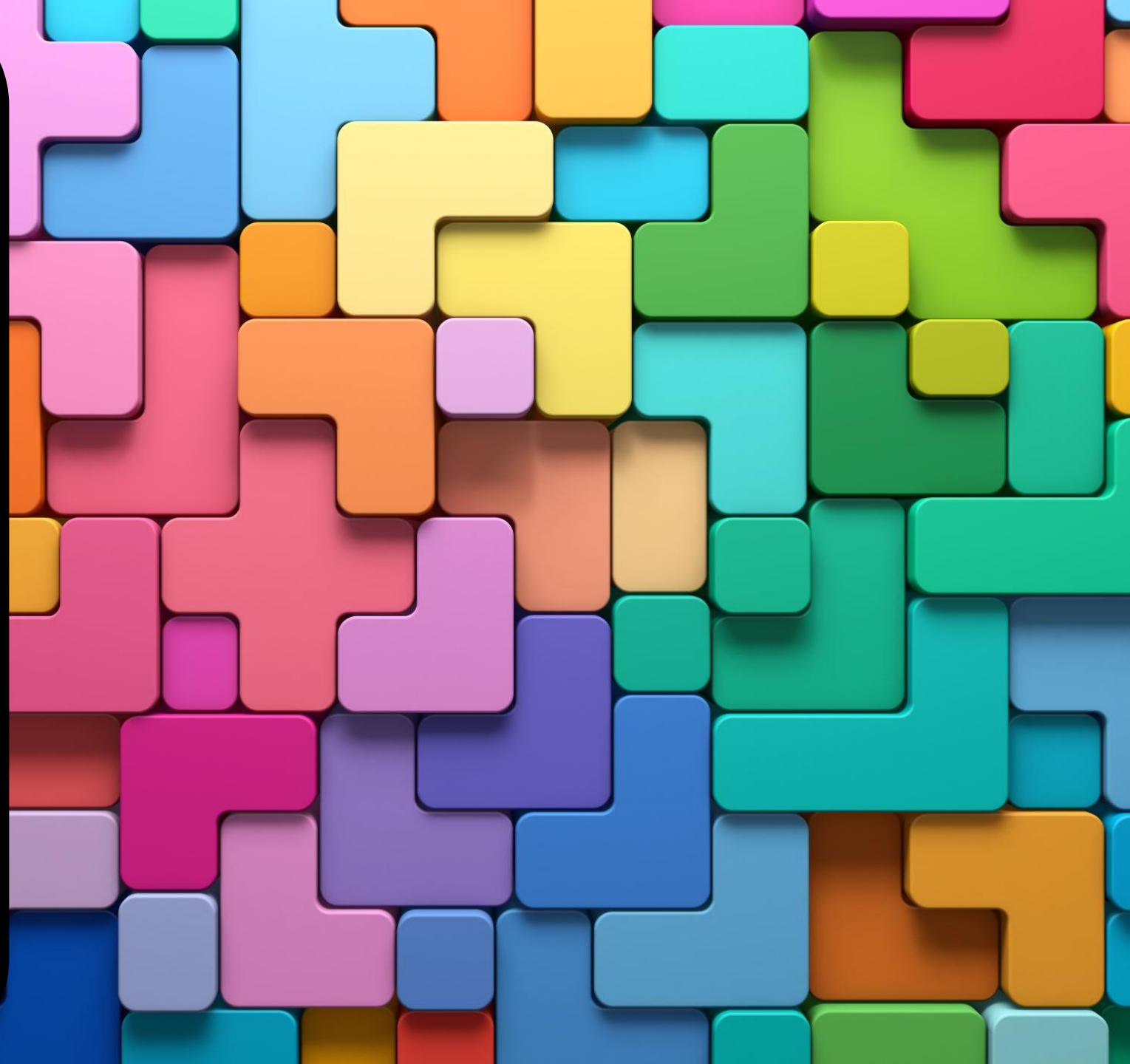
Modeling Early Autism Detection in Children

Team B

Presented by

Kalyan Ghimire, Sandra Lopez Padilla, Clion Muhoza, and Olivier Niyonshuti Mizero



Research Problem

- Autism is often not diagnosed in a timely or efficient manner.
- Early diagnosis is crucial because timely intervention can significantly improve developmental outcomes (Hyman et *al.*, 2020).
- Diagnosing ASD at an early stage remains a challenge due to the wide range of symptom variations and the reliance on expert assessment.

Objectives of the Study

- Develop a predictive model with higher accuracy for early autism detection using machine learning techniques.
- Identify factors that contribute to early autism detection and improve the accuracy of classification models used for ASD screening.

Data Overview (1)

- The data set Autistic Spectrum Disorder Screening Data for Children is introduced by Fadi Fayez Thabtah from the Department of Digital Technology of the Manukau Institute of Technology in Auckland, New Zealand.
- The data set consists of 292 instances and 21 attributes (12 numerical variables and 9 categorical variables).
- The data set recorded 10 behavioral features from the child version of the Autism Spectrum Quotient-10 (AQ-10-Child) screening test and 10 ASD related characteristics.
- The variable names are: 'A1_Score', 'A2_Score', 'A3_Score', 'A4_Score', 'A5_Score', 'A6_Score',
- 'A7_Score', 'A8_Score', 'A9_Score', 'A10_Score', 'age', 'gender',
- 'ethnicity', 'jaundice', 'autism', 'country_of_res', 'used_app_before',
- 'result', 'age_desc', 'relation', 'class'
- **Note: *"autism"*** variable means whether any immediate family member has a pervasive developmental disorder (PDD) (yes or no).
- The target variable is 'class' (whether a child has ASD traits YES or NO).

Data Overview (2)

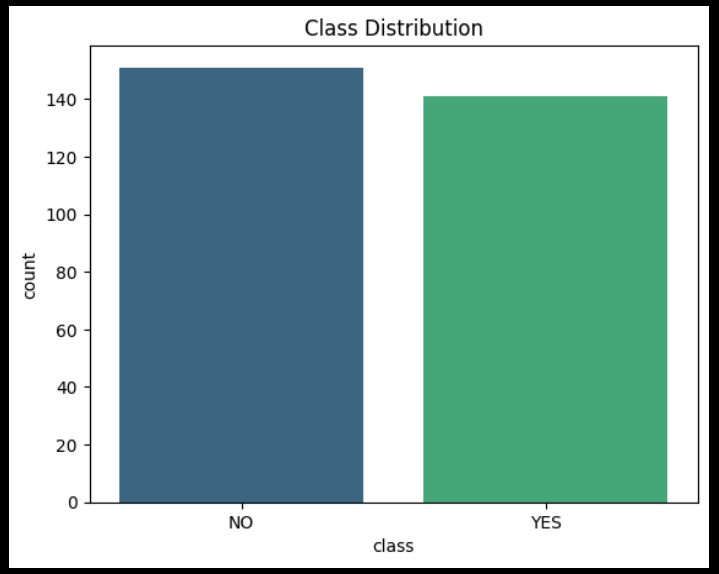
AQ-10-Child screening test consists of the following questions:

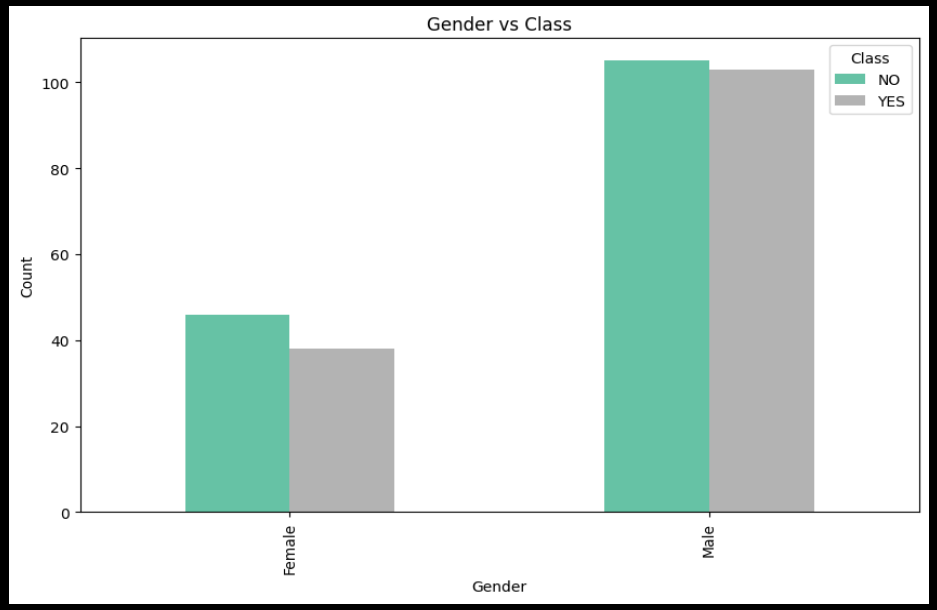
- A1_Score: S/he often notices small sounds when others do not
- A2_Score: S/he usually concentrates more on the whole picture, rather than the small details
- A3_Score: In a social group, s/he can easily keep track of several different people's conversations
- A4_Score: S/he finds it easy to go back and forth between different activities
- A5_Score: S/he doesn't know how to keep a conversation going with his/her peers
- A6_Score: S/he is good at social chit-chat
- A7_Score: When s/he is read a story, s/he finds it difficult to work out the character's intentions or feelings
- A8_Score: When s/he was in preschool, s/he used to enjoy playing games involving pretending with other children
- A9_Score: S/he finds it easy to work out what someone is thinking or feeling just by looking at their face
- A10_Score: S/he finds it hard to make new friends
- result: sum of the 10 scores.

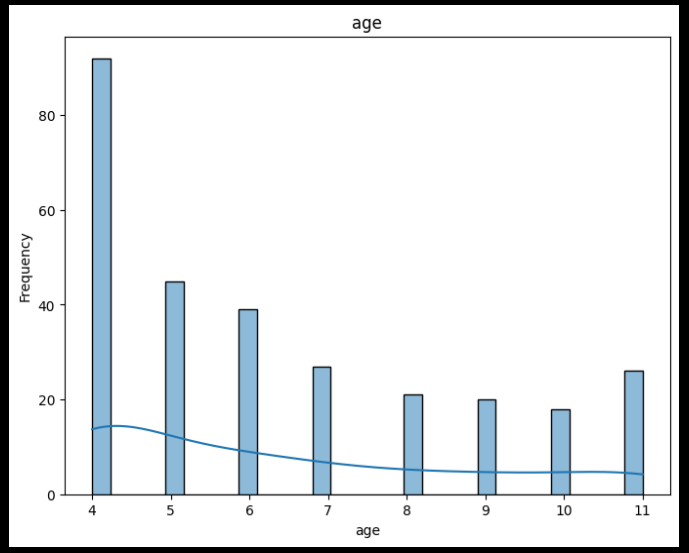
Data Cleaning & Preprocessing

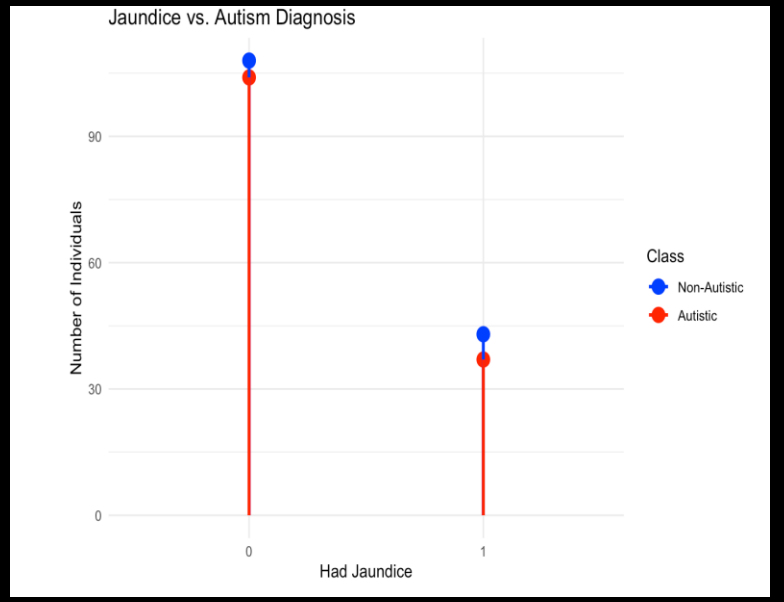
1. Handling Missing data:
   - ***"ethnicity"*** (**43 missing**) and ***"relation"*** (**43 missing**) were filled with ***"Unknown"***.
   - ***"age"*** (**4 missing**) was imputed using the median age to maintain consistency.
2. Fixing Data Issues:
   - Removed unnecessary columns: ***"result"*** and ***"age_desc"*** (not useful for analysis).
   - Standardized categorical values to fix inconsistencies.
3. Encoding Data for Models:
   - Converted binary categories (e.g., ***"gender"***, ***"autism"***) using ***Label Encoding*** (0/1).
   - Used ***One-Hot Encoding*** for multi-category features like ***"ethnicity"***.
4. Scaling & Standardization:
   - Applied ***Min-Max Scaling*** to numerical features to bring all values to a common scale.

Univariate & Multivariate Analysis (1)

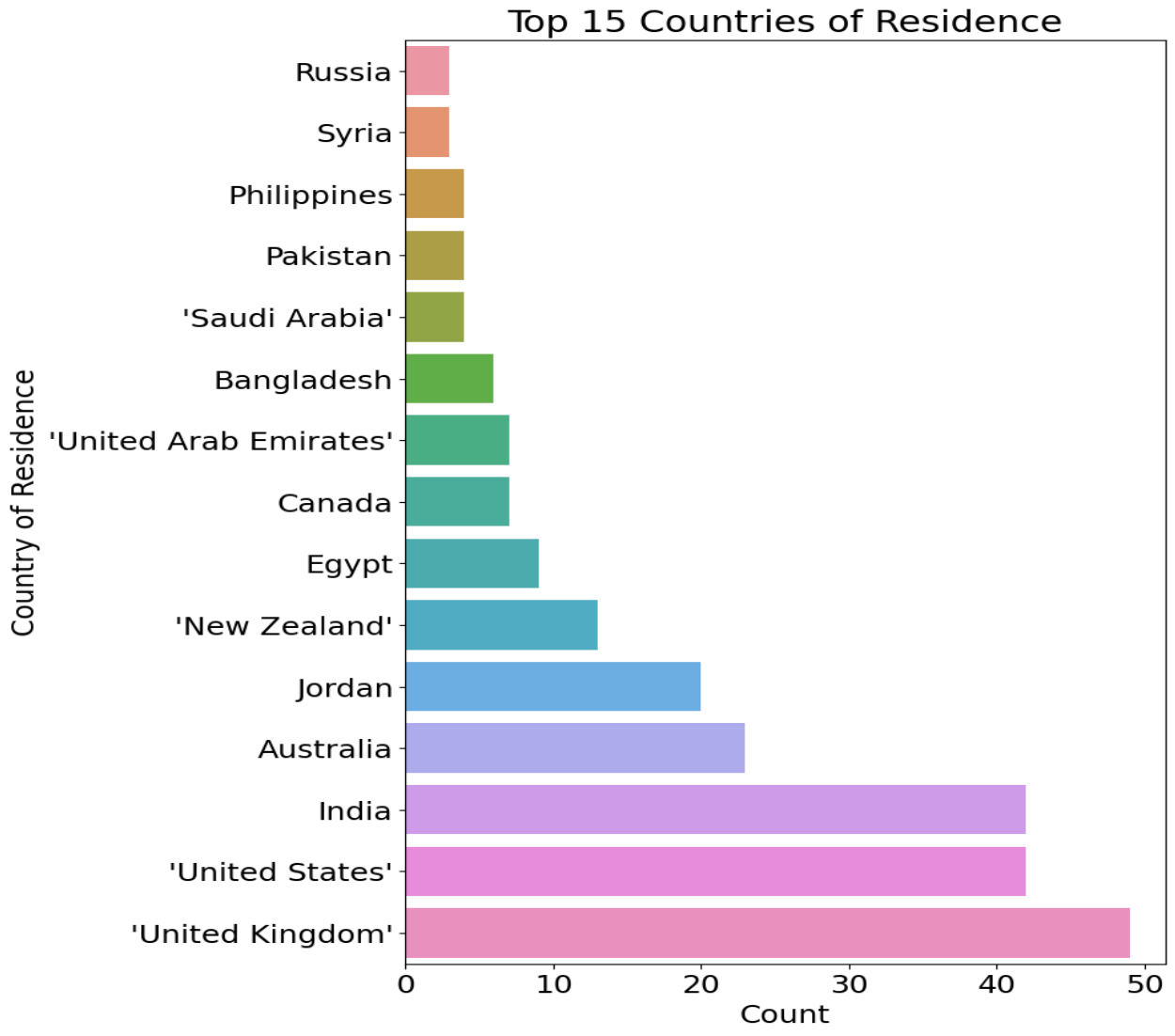


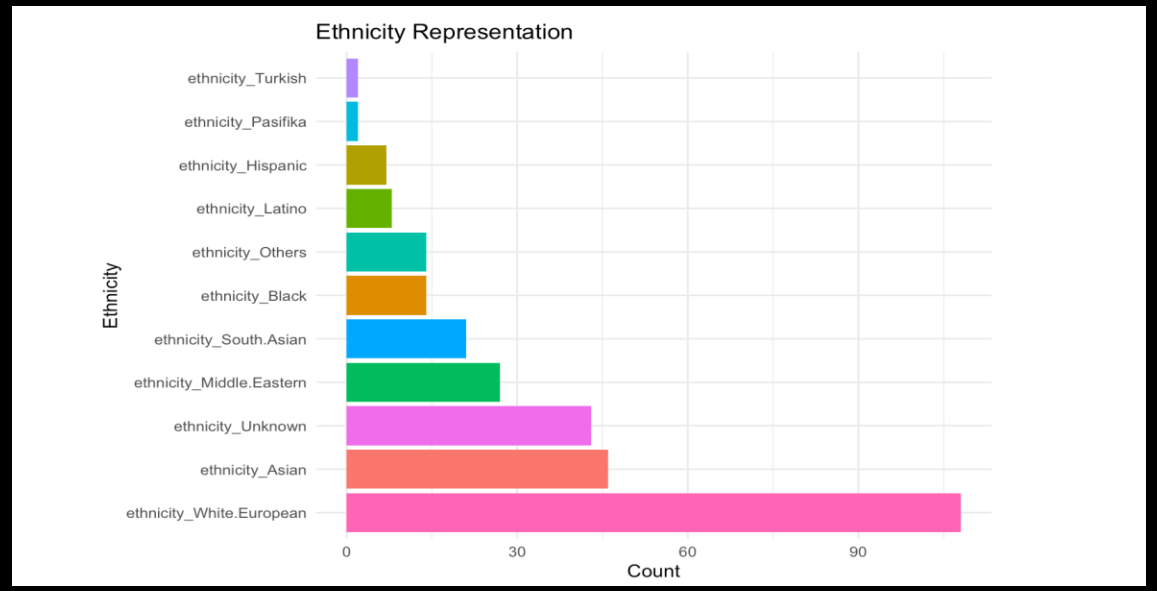


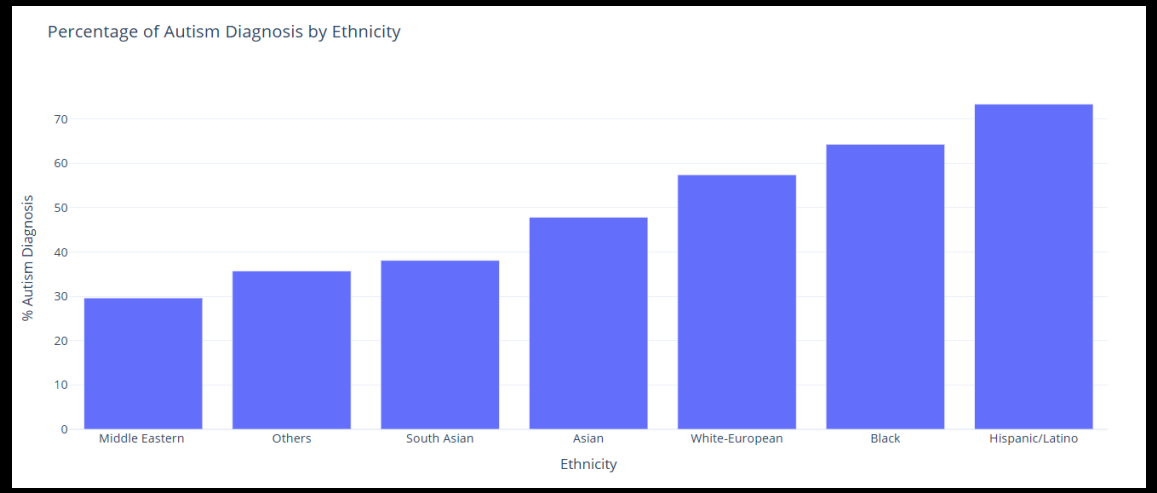




Univariate & Multivariate Analysis (2)







Limitations & Challenges

1. Small Dataset Size:
   - Only ***292*** records, which may not fully represent all autism screening cases.
2. Self-Reported Screening vs. Clinical Diagnosis:
   - The dataset is based on a questionnaire filled out by parents/caregivers, not formal clinical assessments which can be subjective and may introduce bias or misinterpretation.
3. Limited Medical & Genetic Data:
   - While the dataset includes family autism history and jaundice, it lacks more medical history or clinical evaluations.
4. Missing data Challenges:
   - 43 missing values in ***"ethnicity"*** and ***"relation"***, requring imputation.
   - 4 missing values in ***"age"***, handled using the median.

Next Steps...

1. Feature Selection & Optimization:
   - Analyze the importance of each feature and remove those that add little value.
2. Improving Model Performance:
   - Test different machine learning models (***Logistic Regression***, ***XGBoost***, ***Decision Trees***, ***Random Forest***).
   - Tune hyperparameters to improve accuracy and reduce overfitting.
3. Cross-Validation & Testing:
   - Use train-test splits and k-fold-cross-validation to ensure reliable results.
   - Compare model accuracy using metrics like ***precision***, ***recall***, and ***F1-score***.
4. Final Report & Interpretation:
   - Summarize key insights in a clear and simple way.
   - Highlight findings that could help improve early autism detection.
5.

Thank you!