MATH 4323 Part II: Data Selection

Group members: Adriana Alvarez, Emily Gamez, Lendy Varela, Grishma Vemireddy

Data: Autism-Adult-Data.arff / Autistic Spectrum Disorder Screening Data for Adult Dataset

Source: This dataset came from UCI Machine Learning Repository, Autism Screening Adult - UCI Machine Learning Repository

Inspiration behind the dataset: Autistic Spectrum Disorder (ASD) is a neurodevelopmental condition associated with costly healthcare expenditures. Waiting times for a diagnosis are lengthy and procedures are not cost effective for patients. The increase in the number of ASD cases shows that there is a need for easily implemented and effective screening methods that could help health professionals as well as individuals in deciding whether they should pursue a clinical diagnosis.

Size of the data (# of observations and # of variables): 704 observations, and 21 predictors including the class label.

Description of all variables:

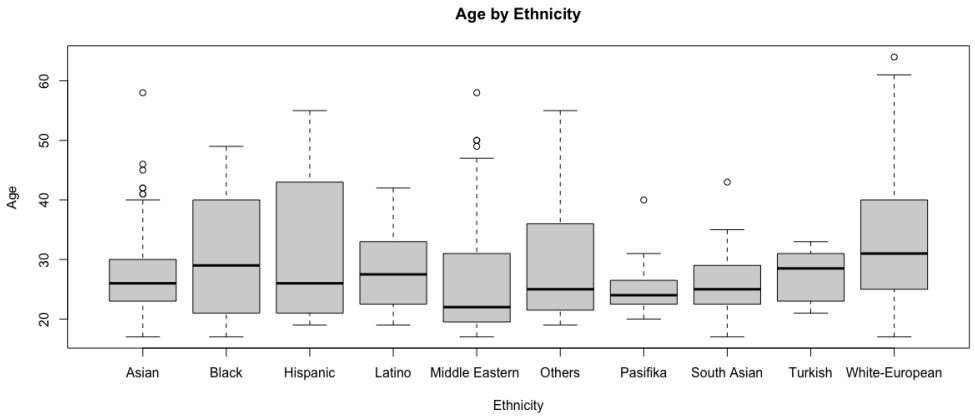
- A1_score (categorical) - Score to answer 1 (Binary 0 – 1)
- A2_score (categorical) - Score to answer 2 (Binary 0 – 1)
- A3_score (categorical) - Score to answer 3 (Binary 0 – 1)
- A4_score (categorical) - Score to answer 4 (Binary 0 – 1)
- A5_score (categorical) - Score to answer 5 (Binary 0 – 1)
- A6_score (categorical) - Score to answer 6 (Binary 0 – 1)
- A7_score (categorical) - Score to answer 7 (Binary 0 – 1)
- A8_score (categorical) - Score to answer 8 (Binary 0 – 1)
- A9_score (categorical) - Score to answer 9 (Binary 0 – 1)
- A10_score (categorical) - Score to answer 10 (Binary 0 – 1)
- age (numeric) - Age in years
- gender (categorical) - Male or Female
- ethnicity (categorical) - List of common ethnicities
- jundice (categorical) - Whether the case was born with jaundice
- country_of_res (categorical) - List of countries (country of residence)
- used_app_before (categorical) - Whether the user has used a screening app
- result (numeric) – Screening score
- age_desc (categorical) - Age description

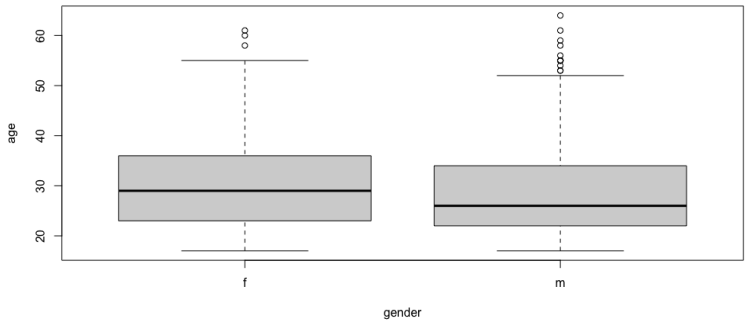
- relation (categorical) - Parent, self, caregiver, medical staff, clinician
- Class/ASD (categorical) - Whether the app classified them with autism or not
- autism (categorical) - Whether the person was diagnosed with autism

Data cleaning:

- 95 observations with missing data, a total of 192 missing data cells
- Only 2 NA in age (numerical): observations 63 and 92
- 95 NA in ethnicity (categorical) and 95 NA in relation (categorical)– both are missing in the observations.
- 13.5% of observations have NA values.
- Observation 53 has age = 383, not possible
- For ethnicity, there are two categories: “others” and “Others”. Combine both together.

Plots of the diversity of the samples after data cleaning.





Citation: Thabtah, F. (2017). Autism Screening Adult [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5F019.