Project Proposal

The Repos- Jeffrey Bohrer, Alexandra Green, Anna Zhang, Kevin Lee

```r
library(tidyverse)
library(tidymodels)
library(foreign)

# load data
autism_ds <- read.arff("Autism-Adult-Data.arff")
autism_ds <- autism_ds %>%
  rename(small_sounds = A1_Score,
         difficult_to_understand_char  = A2_Score,
         ease_to_read_between_lines = A3_Score ,
         focus_on_whole_picture = A4_Score,
         i_can_tell_if_someone_bored = A5_Score,
         i_can_multi_task = A6_Score,
         i_can_tell_feelings_from_faces = A7_Score,
         i_can_go_back_to_work_when_interrupted = A8_Score,
         i_like_to_collect_info_on_categories = A9_Score,
         i_find_it_hard_to_figure_out_others_intentions = A10_Score)
autism_ds <- na.omit(autism_ds)
##Source:
## Based on AQ-10 Test:https://embrace-autism.com/aq-10/
## Score of 6 or higher is indicative of autistic traits
##UCI Machine Learning Repository. Autism Screening Adult Data Set.
##University of California, Irvine, 2018,
##https://archive.ics.uci.edu/dataset/426/autism+screening+adult.
```

Introduction

...

Data description

Our data set on the autism screening of adults is published in the UCI Machine Learning Repository. It was created by Prof. Fadi Fayez Thabtah of the Manukau Institute of Technology in New Zealand in 2017. The data was sourced from voluntary participants of his app, ASDTests, which screens its users for potential indicators of autism using a 10-question survey. There are 704 observations in the original data set and 20 different features: 10 being the answers to the questions of the survey about certain behaviors related to autism, and the remaining 10 being characteristics of the individual, like demographics of age and gender. We can determine the presence of autism to be likely and requiring further testing if an individual receives a score of 6 or higher, as found by Prof. Thabtah's research.

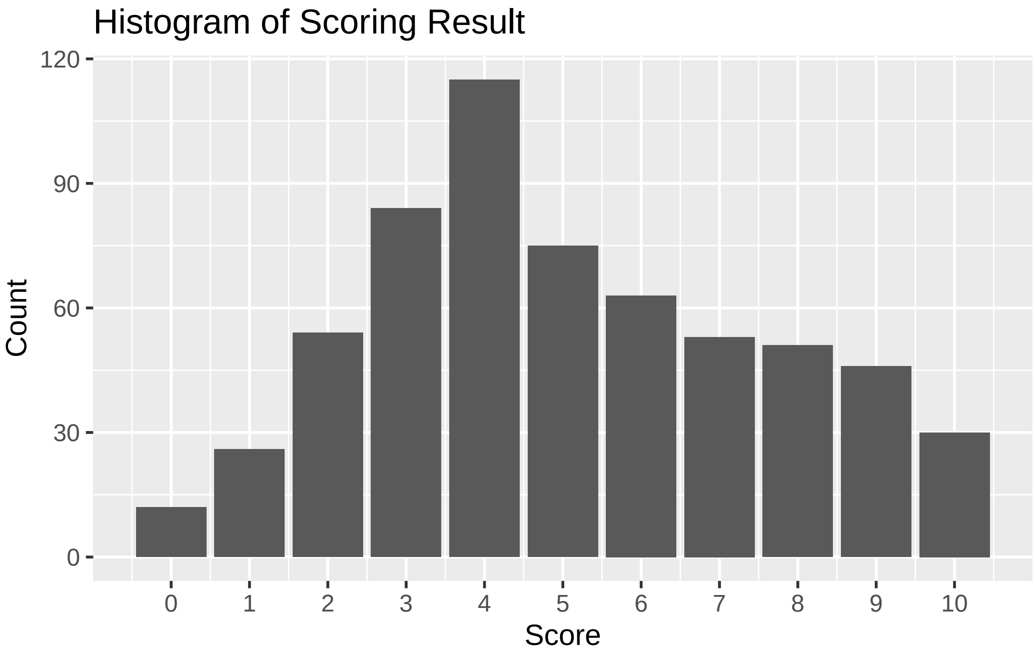
Exploratory data analysis

Since the dataset from the website is in `.arff` (Weka Attribute-Relation File Format), the first step of processing the data is importing the `.arff` into R as a usable dataframe. The function to do so is `read.arff`, and it is stored as `autism_ds`. In the dataframe, some columns are named with abbreviated variable names, such as `A1_Score`, `A2_Score`, etc. These are not very descriptive, and it might cause confusion in analyzing and interpreting the data. Therefore, we renamed them to be more specific, based on the corresponding questions in the AQ-10 test. For instance, the first question is about noticing small sounds when others do not, so we renamed `A1_Score` to `small_sounds` in the dataframe. We renamed 10 such variables in total. The original dataset contains missing values. Having NAs when fitting regression models might affect the models' performance, so we used `na.omit` to remove observations with at least one missing values. The updated dataframe has 609 observations.

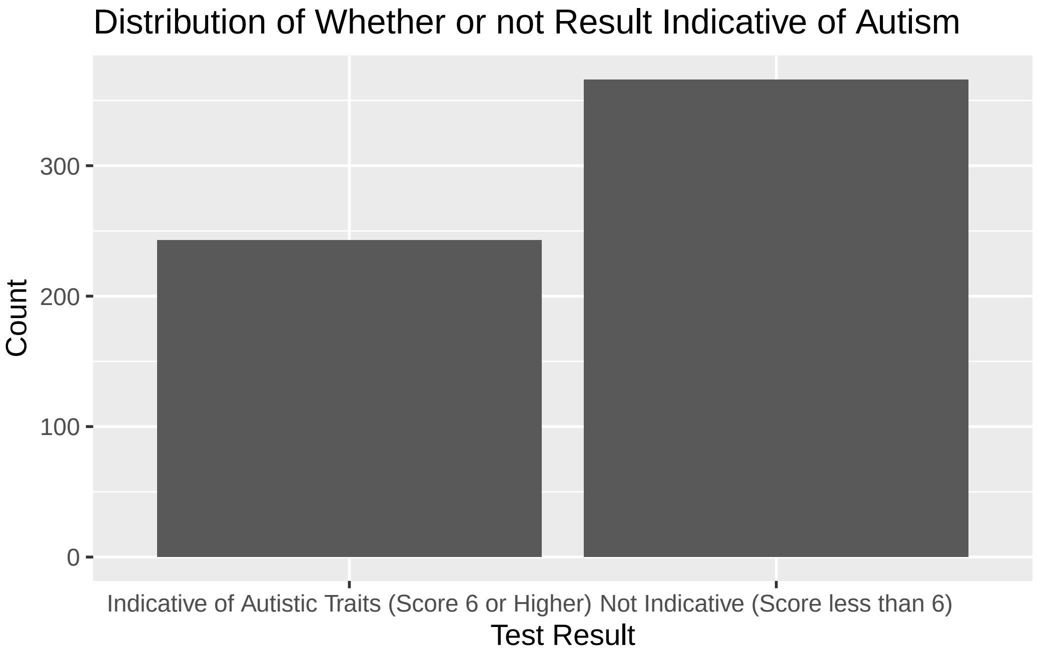
```r
summary(autism_ds$result)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   3.000   5.000   5.077   7.000  10.000 
```

```r
autism_ds |>
  ggplot(aes(x = result)) +
  geom_bar() +
  scale_x_continuous(breaks = seq(0, 10, 1),
                     labels = as.character(seq(0, 10, 1))) +
  labs(x = "Score",
       y = "Count",
       title = "Histogram of Scoring Result")
```



```r
autism_ds |>
  mutate(result = ifelse(autism_ds$result >= 6,
                "Indicative of Autistic Traits (Score 6 or Higher)",
                "Not Indicative (Score less than 6)")) |>
  ggplot(aes(x = result)) +
  geom_bar() +
  labs(x = "Test Result",
       y = "Count",
       title = "Distribution of Whether or not Result Indicative of Autism")
```



The response variable `result` is a score between 0 and 10 from the AQ-10 questionnaire. According to the questionnaire, if the score is 6 or higher, it suggests that the person might have autism. If the score is less than 6, the person may not be autisitc. The distribution of the response variable `result` is fairly symmetrical with a mode at 4, and the distribution has a center at 5, which is the median. The spread can be measured by the IQR, which is 4 (Q3 - Q1 = 7 - 3 = 4). Based on the visualization, there does not appear any apparent outliers.

Analysis approach

The clear response variable of interest is "Class/ASD," given that it is the categorical measure of whether a test-taker possesses enough traits to be deemed in the spectrum. Considering the binary nature of the variable—taking values "Yes" or "No"—it is plausible to postulate that a logistic regression model is an optimal model for its prediction.

In terms of relevant predictor variables, the most interesting ones to add to such a model would be individual AQ-10 question responses (binary, categorical)—given that these are the main questions used to provide a final diagnosis. Additionally, considering the interaction terms between such variables along with sex, age, and ethnic background would be interesting due to the prevalence of mentions in the literature that men tend to be diagnosed more often than women (Brickhill et al., 2023), older folk tend to be underdiagnosed and under-researched (Mason et al., 2022), and that white patients are more likely to be diagnosed.

As a means of assessing such an ambitious research question, multiple models will be fitted until arriving the most parsimonious one and assessing whether they have a statistically significant effect on the final diagnosis as provided by the app.

In terms of relevant predictor variables, the most interesting ones to add to such a model would be individual AQ-10 question responses (binary, categorical)—given that these are the main questions used to provide a final diagnosis. Additionally, considering the interaction terms between such variables along with sex, age, and ethnic background would be interesting due to the prevalence of mentions in the literature that men tend to be diagnosed more often than women (Brickhill et al., 2023), older folk tend to be underdiagnosed and under-researched (Mason et al., 2022), and that white patients are more likely to be diagnosed. As a means of assessing such an ambitious research question, multiple models will be fitted until arriving the most parsimonious one and assessing whether they have a statistically significant effect on the final diagnosis as provided by the app.

Data dictionary

The data dictionary can be found here.

References

Brickhill, Rae et al. "Autism, thy name is man: Exploring implicit and explicit gender bias in autism perceptions." PloS one vol. 18,8 e0284013. 23 Aug. 2023, doi:10.1371/journal.pone.0284013

Mason, David et al. "Older Age Autism Research: A Rapidly Growing Field, but Still a Long Way to Go." Autism in adulthood : challenges and management vol. 4,2 (2022): 164-172. doi:10.1089/aut.2021.0041

Aylward, Brandon S et al. "Racial, Ethnic, and Sociodemographic Disparities in Diagnosis of Children with Autism Spectrum Disorder." Journal of developmental and behavioral pediatrics : JDBP vol. 42,8 (2021): 682-689. doi:10.1097/DBP.0000000000000996

Thabtah, Fadi, et al. "A New Computational Intelligence Approach to Detect Autistic Features for Autism Screening." Applied Soft Computing, vol. 83, 2019, 105748. ScienceDirect, https://doi.org/10.1016/j.asoc.2018.06.011.