EDLD654 ML Project

Maiko Hata

```
# "Beautiful presentations with R and Quarto" https://youtu.be/01KifhHDkFk?si=2axQMI_c0Tu9_Z5
# quarto themes  https://quarto.org/docs/presentations/revealjs/themes.html
# theme - serif or simple
```

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(finalfit)
library(lubridate)
```

```
Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
library(reticulate)
library(finalfit)
library(stringr)
library(recipes)
```

```
Attaching package: 'recipes'

The following object is masked from 'package:stringr':

    fixed

The following object is masked from 'package:stats':

    step
```

```
library(ggplot2)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats 1.0.1     v tibble  3.3.0
v purrr   1.1.0     v tidyr   1.3.1
v readr   2.1.5

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter()  masks stats::filter()
x recipes::fixed() masks stringr::fixed()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(gt)
library(caret)
```

```
Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

    lift
```

```
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```
library(AppliedPredictiveModeling)
```

Research problem

The big question: How can I design and apply Machine Learning (ML) models without reinforcing existing biases?

Asian MIT Student Asks AI for a Pro Headshot, Gets Turned White

AUG 03, 2023 MATT GROWCOOT

Rona Wang, left, and the AI headshot, right, which turned her Caucasian.

Research question

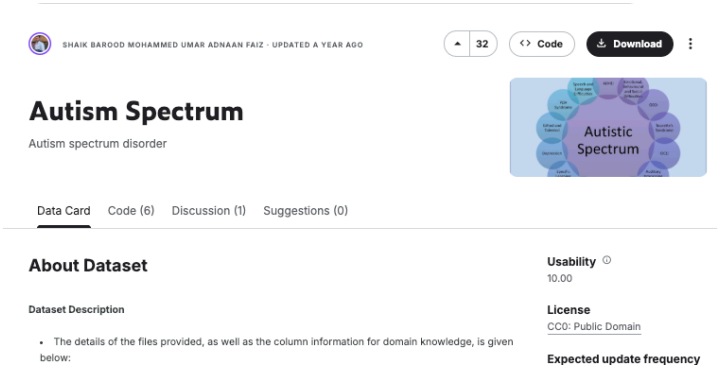
- **Goal**: Predict Autism Spectrum Quotient (AQ-10) screener scores from demographic information.
- **Potential benefit**: Understanding factors relating to scores can potentially support more focused outreach.

- **Ethical & equity considerations**: As an Autistic researcher and Early Intervention (EI) specialist, I want to understand how predictive models are created, what their limitations are, and the ethical considerations.

```
# Machine learning is a topic that holds many layers and emotions. There are clear benefits,
```

Data

- From Kaggle
  - **Numerical** (age, screener result)
  - **Categorical** (gender f/m, jaundice yes/no, family history of Autism yes/no, used app before yes/no, results from screener results YES (7+) NO (~6), ethnicity, country of residence)




```
#I looked for dataset on Kaggle as recommended by Lina. There were a few Autism related datas
```

Cleaning up missing data

```
autism <- read.csv('/Users/maiko/Desktop/EDLD654_ML/data/Autism_Screening.csv', header=TRUE)
```

```
setwd("~/Desktop/EDLD654_ML")
```

```
autism$gender_binary <- ifelse(autism$gender == "m", 1, 0)
autism$jundice_binary <- ifelse(autism$jundice == "yes", 1, 0)
autism$Class.ASD_binary <- ifelse(autism$Class.ASD == "YES", 1, 0)
```

```
ff_glimpse(autism[, 1:20])
```

```
$Continuous
              label var_type   n missing_n missing_percent mean  sd min
A1_Score   A1_Score    <int> 704         0             0.0  0.7 0.4 0.0
A2_Score   A2_Score    <int> 704         0             0.0  0.5 0.5 0.0
A3_Score   A3_Score    <int> 704         0             0.0  0.5 0.5 0.0
A4_Score   A4_Score    <int> 704         0             0.0  0.5 0.5 0.0
A5_Score   A5_Score    <int> 704         0             0.0  0.5 0.5 0.0
A6_Score   A6_Score    <int> 704         0             0.0  0.3 0.5 0.0
A7_Score   A7_Score    <int> 704         0             0.0  0.4 0.5 0.0
A8_Score   A8_Score    <int> 704         0             0.0  0.6 0.5 0.0
A9_Score   A9_Score    <int> 704         0             0.0  0.3 0.5 0.0
A10_Score A10_Score    <int> 704         0             0.0  0.6 0.5 0.0
result       result    <int> 704         0             0.0  4.9 2.5 0.0
          quartile_25 median quartile_75  max
A1_Score          0.0    1.0         1.0  1.0
A2_Score          0.0    0.0         1.0  1.0
A3_Score          0.0    0.0         1.0  1.0
A4_Score          0.0    0.0         1.0  1.0
A5_Score          0.0    0.0         1.0  1.0
A6_Score          0.0    0.0         1.0  1.0
A7_Score          0.0    0.0         1.0  1.0
A8_Score          0.0    1.0         1.0  1.0
A9_Score          0.0    0.0         1.0  1.0
A10_Score         0.0    1.0         1.0  1.0
result            3.0    4.0         7.0 10.0

$Categorical
                          label var_type   n missing_n missing_percent levels_n
age                         age    <chr> 704         0             0.0       47
gender                   gender    <chr> 704         0             0.0        2
ethnicity             ethnicity    <chr> 704         0             0.0       12
jundice                 jundice    <chr> 704         0             0.0        2
austim                   austim    <chr> 704         0             0.0        2
contry_of_res     contry_of_res    <chr> 704         0             0.0       67
used_app_before used_app_before    <chr> 704         0             0.0        2
age_desc               age_desc    <chr> 704         0             0.0        1
relation               relation    <chr> 704         0             0.0        6
               levels levels_count levels_percent
age                 -            -              -
gender              -            -              -
```

```
ethnicity            -            -            -
jundice              -            -            -
austim               -            -            -
contry_of_res        -            -            -
used_app_before      -            -            -
age_desc             -            -            -
relation             -            -            -
```

Cleaning up missing data

- Because the dataset had "?" for missing data, I calculated the percentage of "?" values in the Ethnicity column.

```
sum(autism$ethnicity == "?", na.rm = TRUE)
```

```
[1] 95
```

```
nrow(autism)
```

```
[1] 704
```

```
mean(autism$ethnicity == "?", na.rm = TRUE) * 100
```

```
[1] 13.49432
```

Descriptive statistics

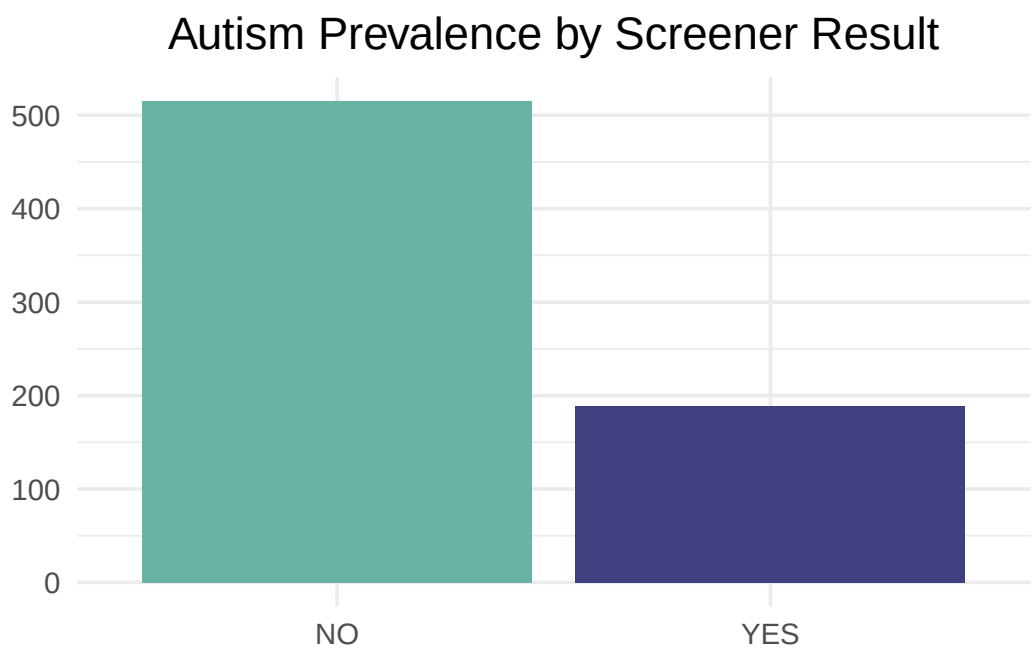
Autism Screener result

```
autism <- autism %>%
  mutate(
    contry_of_res = str_trim(contry_of_res),
    contry_of_res = str_replace_all(contry_of_res, "^[\"']+|[\"']+$", ""),
    contry_of_res = str_squish(contry_of_res),
    contry_of_res = recode(contry_of_res,
      "USA" = "United States", "U.S." = "United States",
      "United States of America" = "United States"
    )
  )
```

```r
autism <- autism %>%
  mutate(
    ethnicity = str_trim(ethnicity),
    ethnicity = str_replace_all(ethnicity, "^[\"']+|[\"']+$", ""),
    ethnicity = str_squish(ethnicity)
    )
```

```
Warning: NAs introduced by coercion
```

```r
ggplot(autism, aes(x = Class.ASD, fill = Class.ASD)) +
  geom_bar() +
  scale_fill_manual(values = c("NO" = "#69b3a2", "YES" = "#404080")) +
  labs(title = "Autism Prevalence by Screener Result",
       x = NULL, y = NULL) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5))
```



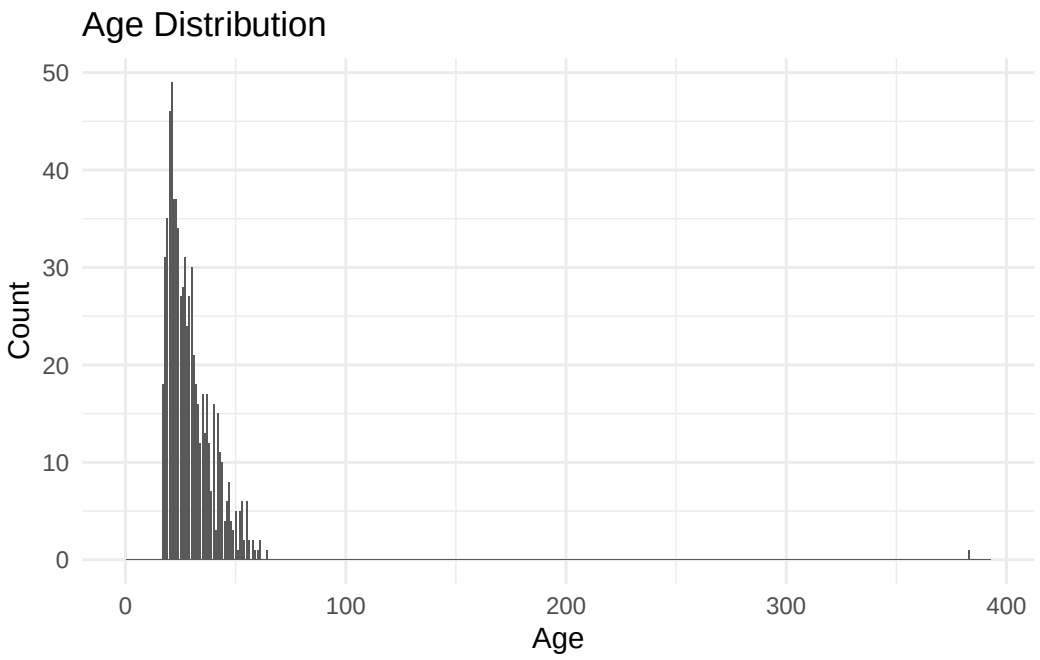
Descriptive statistics

Autism Screener user age

```r
ggplot(autism, aes(x = age)) +
  geom_histogram(binwidth = 0.5) +
  scale_x_continuous(limits = c(0, max(autism$age, na.rm = TRUE) + 10)) +
  labs(title = "Age Distribution",
       x = "Age",
       y = "Count") +
  theme_minimal()
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_bin()`).
```

```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_bar()`).
```

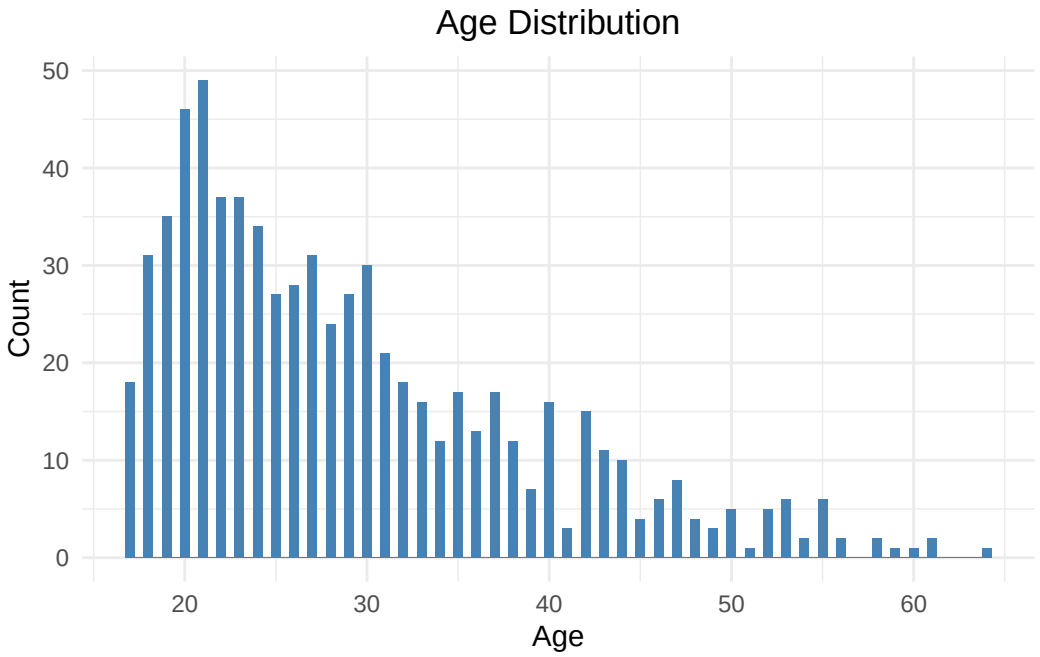


Descriptive statistics

Autism Screener user age with `filter(age <= 100)`

```r
autism <- autism %>% filter(age <= 100)

ggplot(autism, aes(x = age)) +
  geom_histogram(binwidth = 0.5, fill = "steelblue") +
  labs(title = "Age Distribution",
       x = "Age",
       y = "Count") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



Model 1

"… all model building efforts are constrained by the existing data" (p. 61, Kuhn & Johnson, 2016).

Model 2

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Model 3

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Specific setting for model fitting

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Evaluating model performance

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Model fit - Model performance

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Model fit - Final model selection

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Model fit - Cut off point

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Model fit - Other considerations

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Data visualization 1

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Data visualization 2

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Discussion - What I learned

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Discussion - Variables

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Discussion - Findings

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this:

Conclusion

When you click the **Render** button a presentation will be generated that includes both content and the output of embedded code. You can embed code like this: