CS 5821 Final Project:

Exploratory Data Analysis using Machine Learning Techniques

Title: EDA on Autism Screening Data

Team member:

Darryl Lee
Matthew Probosutejo
Thong Sheng Ang

0. Abstract

Autistic Spectrum Disorder, in short, ASD is one of the many disabilities that has no treatment. This disability carries significant healthcare costs to families that are responsible for children that started showing signs of autism at a later time. Thus, the ability to be able to classify whether a kid has autism or not serves a great benefit to healthcare providers and for children itself as well. We managed to find 3 datasets from kaggle that contain autism screening data from both those that have autism and do not. The patients are both from children that are categorized as “below 18” and adults that are categorized as “18 and above”. These datasets are new and used to create a more precise model due to the fact that there exist very little dataset related to patients with autism. We will be applying some data analysis techniques and answering some of our questions and curiosity. Using R Studio, we found out that the data set we used was not enough to represent the entire world of autism patients.

1. Introduction

The purpose of this project is to explore what are the signs that can allow us to deduce whether a person has autism. We will also cover some common myths that have been stated regarding autism.

1.1 Data:

The data we found were from Kaggle. There were 2 sets of data, one for children below the age of 18 and adults that are above 18. The children’s data is split between 2017 and 2018 which has a total of 801 rows while the adult’s data has 609 rows. Both of these data have 21 columns each. The column names are identical with a couple of exceptions that some of them contain typing errors from the creator. There were also a couple of irrelevant data for us. A1_Score to A10_Score are questions that were asked during the screening process. It was not stated in the dataset but it was represented with 1 or 0.

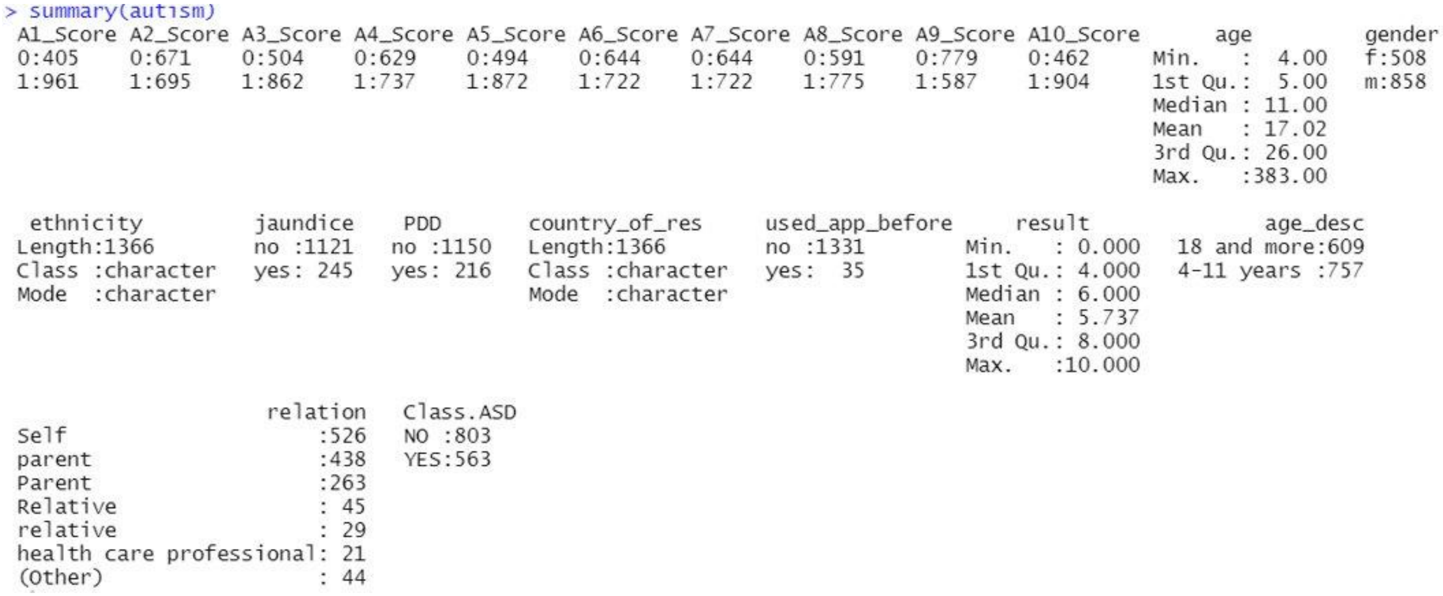
2. Methods and Results

2.1 Data Cleansing

Data cleansing is undeniably one of the most important steps when doing data analysis. Without data cleansing, the results that we obtain might be inaccurate due to errors in the dataset. Five major changes have been made in our data cleansing process, which are:

1. Changing variable classes
2. Removing rows with NA values
3. Removing unused columns
4. Renaming variables
5. Merging datasets into one data frame

The classes of each variable were simply changed using the “colClasses” command when reading in the datasets with the “read.csv” command. Some of the variables were incorrectly labeled in the raw dataset (e.g.: “gender” was labeled as a character variable instead of a factor variable). Next, we remove the rows that contain missing values. It is extremely important to remove any missing data as datasets with missing values tend to yield inaccurate analysis results. There is no such thing as the “best” way when dealing with missing data. Therefore, our team decided to simply remove any records that contain missing values by using the “na.omit” command in R. Unused columns or variables (e.g.: “relation” column) were also filtered out in the process. After that, we renamed some of the columns in the dataset to fix typos as well as to ensure all three datasets share the same column names to facilitate the merging process later on. Lastly, all three datasets are combined into one data frame using the “rbind” function, and a summary of the final data frame can be seen below:



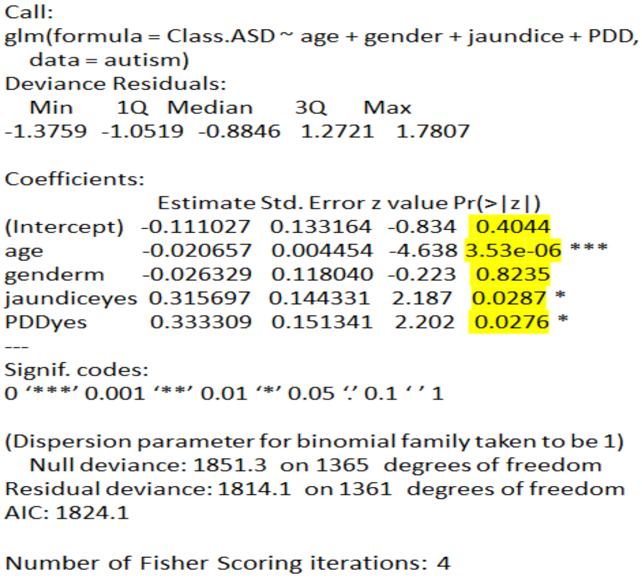
```
> summary(autism)
 A1_Score A2_Score A3_Score A4_Score A5_Score A6_Score A7_Score A8_Score A9_Score A10_Score      age             gender
 0:405    0:671    0:504    0:629    0:494    0:644    0:644    0:591    0:779    0:462     Min.   :  4.00   f:508
 1:961    1:695    1:862    1:737    1:872    1:722    1:722    1:775    1:587    1:904     1st Qu.:  5.00   m:858
                                                                                          Median : 11.00
                                                                                          Mean   : 17.02
                                                                                          3rd Qu.: 26.00
                                                                                          Max.   :383.00

  ethnicity         jaundice    PDD        country_of_res     used_app_before     result               age_desc
 Length:1366        no :1121   no :1150   Length:1366        no :1331        Min.   : 0.000   18 and more:609
 Class :character   yes: 245   yes: 216   Class :character   yes:  35        1st Qu.: 4.000   4-11 years :757
 Mode  :character                         Mode  :character                   Median : 6.000
                                                                             Mean   : 5.737
                                                                             3rd Qu.: 8.000
                                                                             Max.   :10.000

                    relation   Class.ASD
 Self                    :526   NO :803
 parent                  :438   YES:563
 Parent                  :263
 Relative                : 45
 relative                : 29
 health care professional: 21
 (Other)                 : 44
```

2.2 Multiple linear regression

We are using multiple linear regression in R studio to show which variable is significant in predicting the Class.ASD variable which is explained if the person has or does not have autism as the response variable. and the rest of the variable as the explanatory variable logit.fit1 <- glm(Class.ASD ~ ., data=autism, family=binomial), after we assign it to logit.fit1 we just use summary(logit.fit1) to show the result from our multiple logistic regression, however, these result shown an error message Warning messages: 1: glm.fit: algorithm did not converge, 2: glm.fit: fitted probabilities numerically 0 or 1 occurred. The p-value is also inaccurate since all the p-value is = 1. Since including all the variables shows us an error message, instead we are just going to take some of the variables. Now we are just going to choose four variables among the others since these variables might be related to autism. logit.fit2 <- glm(Class.ASD ~ age + gender + jaundice + PDD, data=autism, family=binomial), use the summary(logit.fit2) to show the result :



```
Call:
glm(formula = Class.ASD ~ age + gender + jaundice + PDD, family = binomial,
   data = autism)
Deviance Residuals:
   Min      1Q   Median      3Q      Max
-1.3759  -1.0519  -0.8846   1.2721   1.7807

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.111027  0.133164  -0.834   0.4044
age          -0.020657  0.004454  -4.638 3.53e-06 ***
genderm      -0.026329  0.118040  -0.223   0.8235
jaundiceyes  0.315697  0.144331   2.187   0.0287 *
PDDyes        0.333309  0.151341   2.202   0.0276 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 1851.3  on 1365   degrees of freedom
Residual deviance: 1814.1  on 1361   degrees of freedom
AIC: 1824.1

Number of Fisher Scoring iterations: 4
```

Only the gender variable was not significant to predicting autism since the p-value was > 0.05. With the result from the multiple logistic regression above we can interpret the estimate coefficient. The estimated coefficient for **Age** is negative. It means that as **Age** increases the probability that the person has autism decreases, holding **gender, jaundice, PDD** fixed. The estimated coefficient for **gender** is negative. It means that the probability that the person has autism is smaller if the person is a male, holding **age, jaundice, PDD** fixed. The estimated coefficient for **jaundice** is positive. It means that the probability that the person has autism is higher if the person has the jaundice, holding **age,gender, PDD** fixed. The estimated coefficient for **PDD** is positive. It means that the probability that the person has autism is higher if the person's family member has autism, holding **age, gender, jaundice** fixed.

2.3 Proportion hypothesis test (male vs. female)

Next thing we are using prop tests to prove the research question that Male have higher chances of having autism compared to females. With the null hypothesis female and male have the same chance to have autism, while alternative hypothesis male have a higher chance to have autism. First we going to count how many male have autism and total male in the data set use nrow(autism[autism$gender=="m" & autism$Class.ASD=="YES",]); nrow(autism[autism$gender=="m",]) which show the result are 359 and 858. Same thing to female nrow(autism[autism$gender=="f" &

autism$Class.ASD=="YES",]); nrow(autism[autism$gender=="f",]) and the result are 204 and 508. Finally, use the r function prop.test(x=c(359,858), n=c(858,508), alternative = “greater”, correct = F)

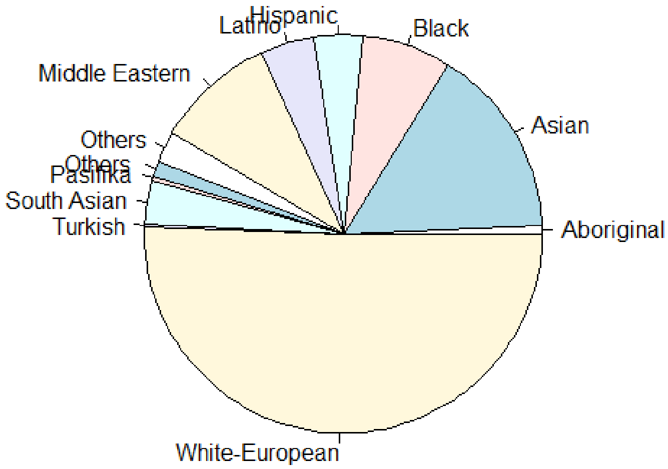
```
2-sample test for equality of proportions without
continuity correction

data:  c(359, 204) out of c(858, 508)
X-squared = 0.37348, df = 1, p-value = 0.2706
alternative hypothesis: greater
95 percent confidence interval:
 -0.02840603  1.00000000
sample estimates:
   prop 1    prop 2
0.4184149 0.4015748
```

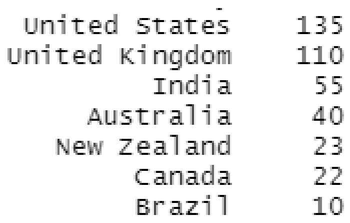
From the p-value = 0.2706 > 0.05 = **α**, we failed to reject the null hypothesis and conclude that there is no evidence to claim that male has higher chance to have autism compare to female.

2.4 Ethnicity and Country of Residents vs Autism Rates

We were trying to see if ethnicity and country of residents have a relation with the rate of autism. To show the distribution of autism among ethnicity, we decided to use a pie chart to represent it. According to our pie chart, it



shows that the majority of the autism patients are white -european, at least from our data set. After researching online, we found the closest data that we can find are the comparison between White, Black and Hispanic from the CDC website. It is said that white children are more likely to be diagnosed with autism compared to black and hispanic. There is no clear relation that White children have a higher rate of autism. The results show otherwise is due to the fact that some families did not have access to healthcare services to have their children diagnosed. Moving forward with the information that we have, we decided to look at it with a bigger picture. Instead of using ethnicity, we decided to look at autism patients in each country. We found that the United States has the highest autism rate, followed by the United Kingdom and India. According to the World Population Review, It is stated that based on 2021, the top countries with highest autism rates are Hong Kong, South Korea and the United States. Looking at our data from ethnicity, it makes sense because the majority of the citizens in both the United States and United Kingdom are White-European. We believe that our data set is not a valid representation for the world. From our data, it is safe to conclude that it does not contain every single autism patient from the world. Majority of the healthcare centers did not participate in contributing to this data, as the creator stated that it is rare to get our hands on data sets with recorded autism patients. We concluded that since both ethnicity and country of residence can’t be proven to be related to the rate of autism, it is inconclusive.



```
 United States    135
United Kingdom    110
         India     55
     Australia     40
   New Zealand     23
        Canada     22
        Brazil     10
```

3. Conclusion

After through all the analysis of the data we now are able to acknowledge what factor may related to autism and among all that we able to prove the ambiguous about autism for example how people say that male has higher chances to have autism compare to woman, however with the analysis from our data we able to show that there’s no different between male and female and both of gender have the same percentage to have autism. Moreover, through our analysis which shows the interpretation of the coefficient on individual variables we are able to see how different factors might increase or decrease the percentage of having autism. With this project I hope in the future this project can go to the next level to help humanity to recognize autism as early as possible with a proper information and not judge everything based on speculation

4. List of References

Centers for Disease Control and Prevention (CDC). (2019, Aug 27). *Spotlight On: Racial and Ethnic Differences in Children Identified with Autism Spectrum Disorder (ASD).* https://www.cdc.gov/ncbddd/autism/addm-community-report/differences-in-children.html

Elshoky, B. (2020, Sep 19). *Autism screening child two version.* [Data set]. Kaggle.com. https://www.kaggle.com/basmarg/autism-screening-child-two-version

Faizunnabi. (2018, Feb 12). *Autism Screening.* [Data set]. Kaggle.com. https://www.kaggle.com/faizunnabi/autism-screening

World Population Review. (2021). *Autism Rates by Country 2021.* https://worldpopulationreview.com/country-rankings/autism-rates-by-country

Zeliadt, N. (2018, June 13). *Autism's sex ratio, explained.* Spectrum. https://www.spectrumnews.org/news/autisms-sex-ratio-explained/