UC DAVIS
COLLEGE OF LETTERS AND SCIENCE
Department of Statistics

Sex Bias in Autism Spectrum Disorder

Allison Peng and Shizhe Chen
Department of Statistics, UC Davis

Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that is characterized by repetitive behaviors, specific interests, or difficulty with social interactions [1].

Once diagnosed, early intervention is strongly advised, however diagnostic methods using behavioral assessments require a specialist evaluation, leading to an average wait of three years [2]. Recent studies attempted to expedite the diagnosis process using published MRI scans to build an Autism classification model [3]. However, the classification model's training requires scans from diagnosed ASD patients and contains a disproportionate ratio of males to females. Such **sex bias** [4] could result in inferior classifications for female subjects.

Objective

We propose to address the sampling bias by calibrating the severity levels from behavioral assessments alongside MRI scans across the sexes. Using the calibration method, we can build a classification model that improves the diagnostic results of female subjects.

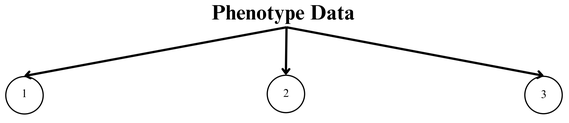
This project focused on a few questions related to the objective:

- Where can we find data that measures both phenotype and MRI data?
- How do we work with a large amount of missing values within a dataset?

ABIDE Autism Brain Imaging Data Exchange → Phenotype Data: Identify Target Variable → Preprocessing: Identify Missing Values → K-Nearest Neighbors to Impute Missing Values → Final Distribution of Target Variable

Data Structure

The **Autism Brain Imaging Data Exchange I** (ABIDE I) is an initiative to combine previously collected resting state functional magnetic resonance imaging (R-fMRI), anatomical and phenotypic datasets from 17 international sites. As a result, the publicly available database consists of 1112 patients, including 539 from individuals with ASD and 573 from typical controls (ages 7-64 years, median 14.7 years across groups). For better understanding, we placed the data into three main groups based on how related they are to the Severity Level (1 is least, 3 is most).



Site ID: Institute of data collection
Patient ID: Unique ID number for each patient
Sex: 1 = Male, 2 = Female
BMI: Body Mass Index at the time of scan
Age at Scan: Age in years at the time of scan
Handedness Category: L = Left, R = Right, Ambi = Ambidextrous
Handedness Score: Measures level at which patient prefers dominant hand
FIQ/PIQ/VIQ Test Type: Type of IQ tests administered
Current Med Status: Is patient currently taking medication?
Medication Name: Active ingredient of any current psychoactive medications, 0 = No, 1 = Yes

Autism Diagnostic Observation Schedule (ADOS)
Module: Categorizes patients into ranking from 1-4
- Module 1: children who are nonverbal or whose vocabulary is limited to a few words
- Module 2: children who use phrase speech
- Module 3: verbally fluent children
- Module 4: verbally fluent adolescents or adults

ADOS Total: Behavioral score from assessment, includes communication and social subscore
ADOS Research Reliable: Was ADOS scored and administered by research reliable personnel? 0 = No, 1 = Yes
Social Communication Questionnaire Total: Behavioral score for initial Autism diagnosis screening

ADOS Gotham Severity Level: Individually calibrated severity score to eliminate influential factors such as age.
ADOS Gotham Total: Total score with Social Affect and Restricted and Repetitive Behaviors assessments
Social Responsiveness Scale: A routinely administered comprehensive diagnostic assessment of ASD
Full Scale Intelligence: Overall score for intelligence
Verbal intelligence: Measured with information test category, digit span, arithmetic, vocabulary, comprehension, similarities test
Performance intelligence: Measured with picture completion, picture arrangement, block design, digit symbol, object assembly

Missing Values

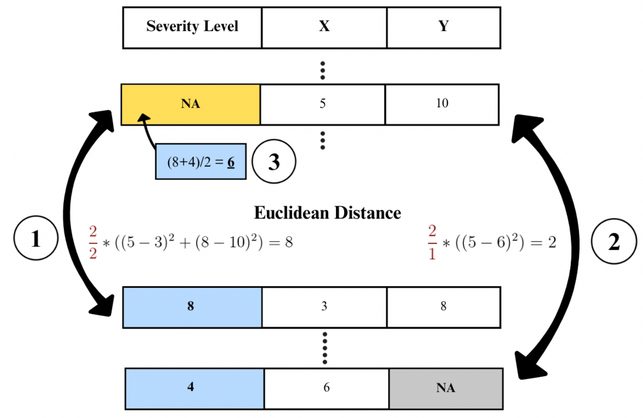
Proportion of columns with missing values in dataset: 92%

Proportion of missing values in dataset: 46.7%

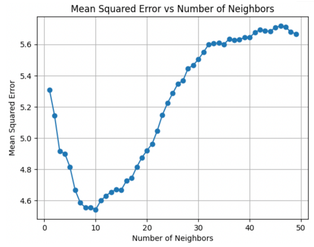
The dataset presented a challenge of containing a large portion of missing data, including missing severity scores. Out of 81 columns, 75 columns contained missing values, or NA values. Although 46.7% of the overall data was missing, we were able to use the remaining 53.3% data to impute the missing values. Prior to imputing, the data was preprocessed by removing unnamed columns, encoding categorical variables and scaled using MinMax scaling.

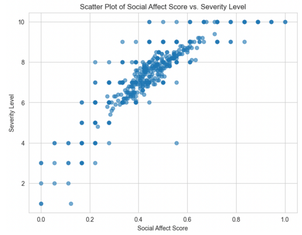
K-Nearest Neighbors

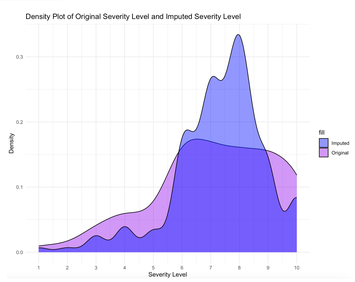
K-Nearest Neighbors (KNN) is a non-parametric supervised learning classifier. The algorithm is also commonly used to impute missing values. KNN imputes a missing data point based on **Euclidean Distance**, and this distance can be modified to handle missing data (as shown by the red numbers). The missing value is imputed based on the majority vote of the closest neighbors.



Results

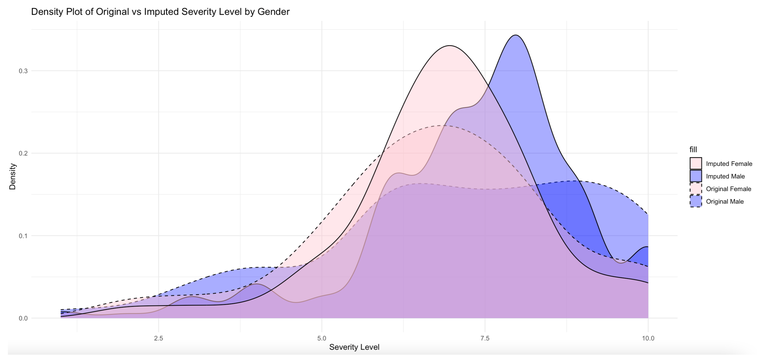






Implemented KNN imputation and calculated accuracy using **K-fold cross validation**. The resulting mean squared error plot showed 10 neighbors as the optimal number of neighbors. Imputing the values allows us to further analyze other features and their relationship with the severity level.

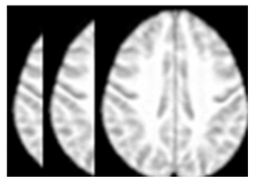
Conclusion



Key Takeaways:

- After imputing, male and female distributions are approximately different from the original distribution with a smaller variance of severity level
- Implementation of KNN algorithm can be used to impute missing data
- Imputed dataset allows us to analyze other variables related to severity score

Next Steps



fMRI Scan [5]

- We aim to analyze the MRI scans to find biomarkers that correlate to severity level
- By using the biomarkers as a standardized way of measuring severity level, we can calibrate the ABIDE severity level to eliminate sex bias for Autism classification models

Acknowledgements & References

I would like to thank Dr. Shizhe Chen for his mentorship and guidance throughout the project.

Resources, e.g., [ABIDE I Dataset] were obtained from www.nitrc.org.

[1] Hodges, H., Fealko, C., & Soares, N. (2020). Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. Translational Pediatrics, 9(S1), S55–S65. https://doi.org/10.21037/tp.2019.09.09
[2] Gordon-Lipkin, E., Foster, J. J., & Peacock, G. (2016). Whittling down the wait time. Pediatric Clinics of North America/the Pediatric Clinics of North America, 63(5), 851–859. https://doi.org/10.1016/j.pcl.2016.06.007
[3] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage. Clinical, 17, 16–23. https://doi.org/10.1016/j.nicl.2017.08.017
[4] Werling, D. M., & Geschwind, D. H. (2013). Sex differences in autism spectrum disorders. Current Opinion in Neurology, 26(2), 146–153. https://doi.org/10.1097/wco.0b013e32835ee548
[5] Lama, R. K., Kim, J., & Kwon, G. (2022). Classification of Alzheimer's disease based on Core-Large Scale Brain Network using multilayer Extreme learning Machine. Mathematics, 10(12), 1967. https://doi.org/10.3390/math10121967

NITRC Neuroimaging Tools & Resources Collaboratory