Predicting ASD Traits with Toddler Datasets

Ryan Yang, Computer Science

Mentor: Dr. David Johannsen

Kellogg Honors College Capstone Project 2025

Abstract

It is known that sometimes, Autism Spectrum Disorder (ASD) traits don't really show up until it's too late. When it's diagnosed at a much later age such as adolescence, it gets more difficult to provide effective treatments and therapy for the child. This research project served as a way to implement language models that could detect early signs of ASD traits. By focusing with an emphasis on a toddler's age from the beginning, certain correlations can be determined to help detect early signs of ASD. The goal was to develop a predictive model based on our initial objective. This model would then help to show that by primarily testing out prediction models with preexisting toddler datasets, a new app can be developed later on that tries to add more traits that can be detected with certain attributes. We opted to use pre-existing datasets from Kaggle, a website that shares a lot of data that can help build AI models. Multiple features of data science and machine learning were used to experiment with these datasets including descriptive statistics and logistic regression.

Origins and Background

Before the age range could be determined for this project, there was a medical research report that served as an inspiration for this project along with my personal experiences dealing with ASD traits. A key point was how the earlier you identify ASD, the earlier you can provide treatments and services. It also mentioned how one of the language models that could assist with predicting those traits was Natural Language Processing (NLP). One way that this model could be tested in healthcare was the use of electronic health records (EHRs). The only setback was that these data records were not publicly accessible. Obtaining this type of data wouldn't be feasible in the given timeframe along with other ethical concerns. Instead, Kaggle datasets were used, which were taken from ASDTests, a mobile app that screens autism by using behavioural tests. After a long discussion about possible age ranges and categories, it was ultimately decided that toddlers would be the best fit, specifically 18-36 months.

Materials and Methods

Preexisting datasets were used from Kaggle in the form of Excel spreadsheets in order to conduct different analyses of the data. In this dataset, the following questions A1-A10 listed in the results table were asked with a yes or no response. Additional attributes were also attached to the patient's data like age, sex, ethnicity, jaundice at birth, and if other family members had a history of ASD. Some of these variables were tested in the data with modeling, but it was determined that they didn't have a significant impact on the correlations between those with ASD and those without ASD. Python Jupyter notebooks were also used to test different analyses of the data. In this dataset, the big key in testing these different attributes was the Q-chat scores, a cumulative score out of 10 given from how many questions answered yes. Yes gave a score of 1, and no gave a score of 0. If the response was Sometime/Rarely/Never, the score was still assigned as 1. Then, if a child scored added more than 3, potential ASD traits were determined.

Key Terms

Logistic Regression - a supervised machine learning algorithm that achieves binary classification tasks by predicting the probability of an outcome, event, or observation.

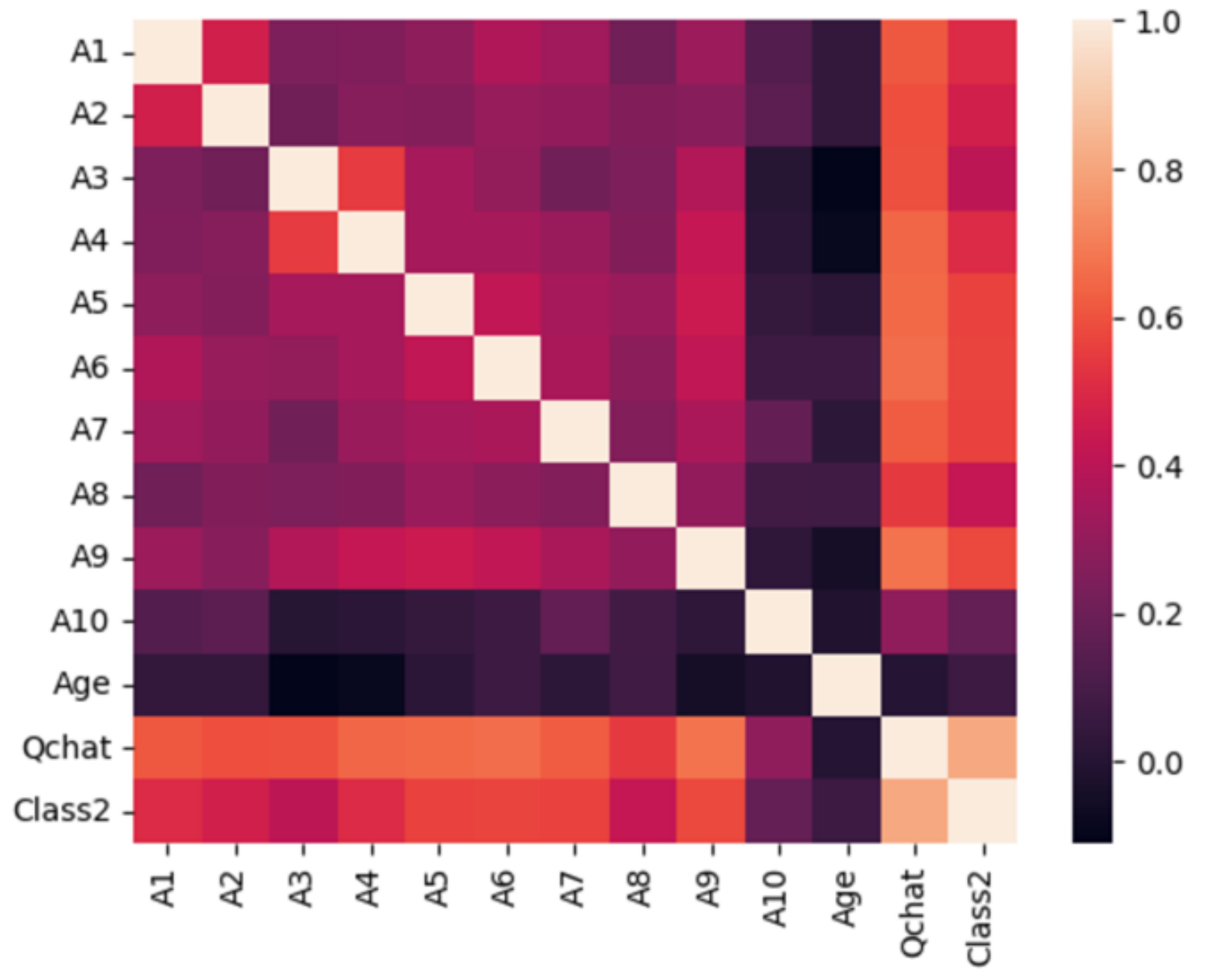
Descriptive Statistics - provides simple summaries about the sample and observations that have been made.

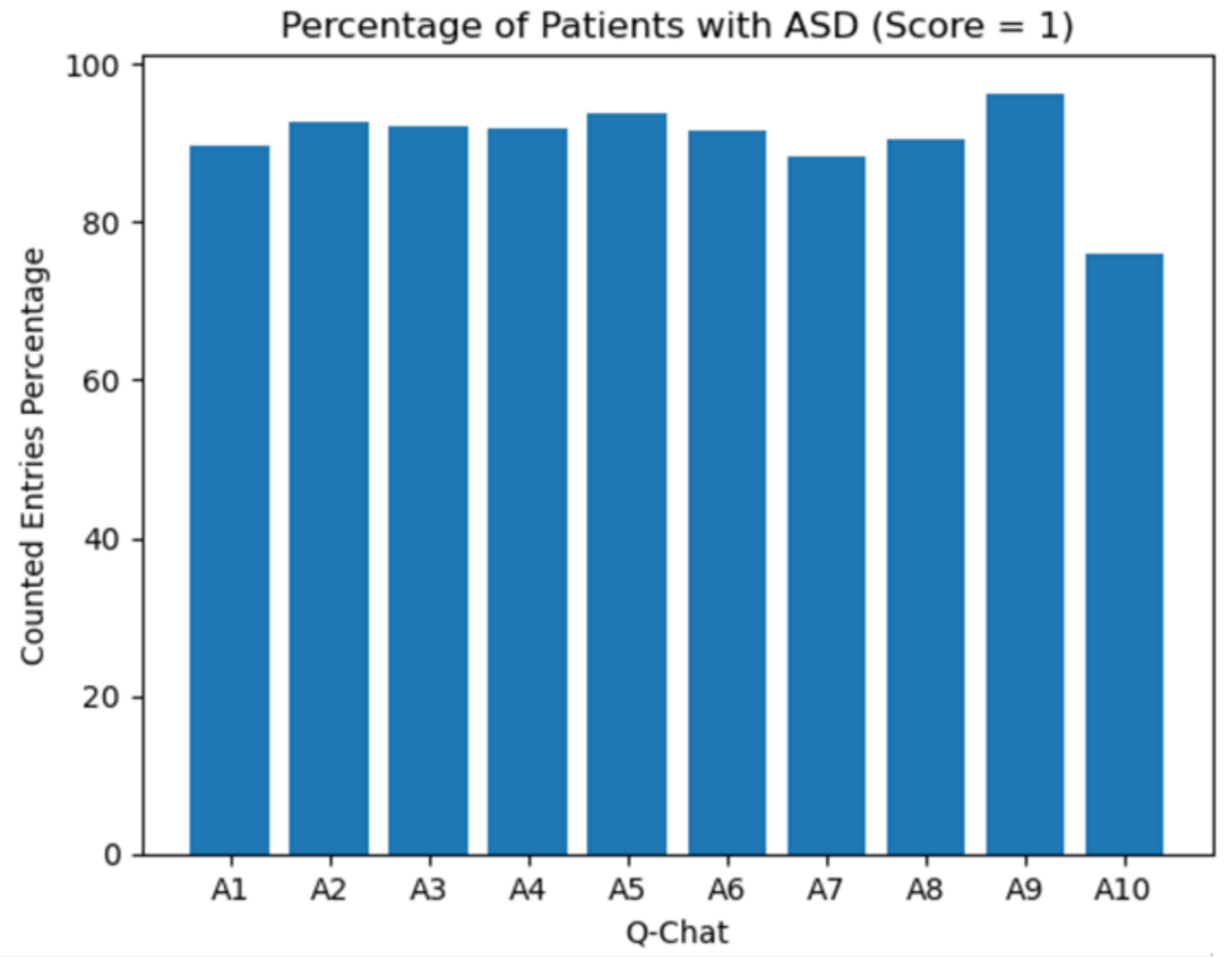
Heatmap - a 2-dimensional data visualization technique that represents the magnitude of individual values within a dataset as a color. Higher color intensity represents stronger correlations.

Results

The results tested for some of the graphs listed below showed some significant correlations. The blue bar graph used descriptive statistics to summarize a collection of patient percentages who scored 1 (or Yes) in the Q-chat. In terms of Q-chat scores, A10 had the lowest percentage in the counted entries, therefore it appeared as the weakest trait. In other words, whether a child stares as nothing won't contribute much to determining ASD traits for toddlers. For A9, it was the highest percentage, so it was determined as the most important correlation with ASD traits. Additional questions like A2, A3, and A5 were the next highest percentages, which also indicate a strong correlation with those who had a score equal to 1.

Some Python code was also implemented to train a machine-learning engine called logistic regression. This algorithm was used to split the data frame into x and y values, scale the values, and split them into testing and training arrays built as sets. These were then tested by using x and y values for each set, which would calculate the accuracy rate at predicting whether a toddler had ASD based on their behavioral screening scores. The final accuracy rate that was calculated for training and test sets were both 100%. Therefore, this would be a highly useful model to predict these traits with high accuracy. The heatmap on the left side was also used as a way to display how strongly each variable can relate to one another with the use of colors. The color bar on the right of the heatmap represents the strength of the correlation: 1.0 being the weakest and 0.0 being the strongest. In fact, the darkest area happened to be around A9, the same question that was deemed the highest percentage on the bar graph.





A1	**Does your child look at you when you call his/her name?**
A2	**How easy is it for you to get eye contact with your child?**
A3	**Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)**
A4	**Does your child point to share interest with you? (e.g. pointing at an interesting sight)**
A5	**Does your child pretend? (e.g. care for dolls, talk on a toy phone)**
A6	**Does your child follow where you're looking?**
A7	**If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g. stroking hair, hugging them)**
A8	**Would you describe your child's first words as:**
A9	**Does your child use simple gestures? (e.g. wave goodbye)**
A10	**Does your child stare at nothing with no apparent purpose?**

Summary and Conclusions

At first, it may seem too good to be true for this data sample to receive 100% accuracy for training and test sets, but this calculation was confirmed and cross-referenced through other data contributors, who all gave the same result. Also, since Question A9 asked about using simple gestures, the strong correlation can indicate a big sign of ASD since those who cannot communicate verbally at an early age may have more trouble in the future when treatment isn't given at an early age. Other attributes in the data chart like gender and jaundice were considered during this project, but they didn't greatly affect how the data was represented. However, one factor that did highly correlate with the data was ethnicity. White European, Asian, and Middle Eastern were shown to have the highest increase of ASD cases.

In summary, this project showcased a stepping stone to developing predictive models which were tested to pretty accurate results. These models were meant to help give others an opportunity with a new mobile/web application to help test toddlers using these datasets in order to detect ASD traits at a young age and help them get the treatment that they need as soon as possible.

Future Works

Future works would include adding additional Kaggle datasets besides the ones tested here. Other works could also add additional predictive models with this data that would lead to the development of a mobile app that can help determine these ASD traits early on. Since these predictive models were implemented by using datasets taken from the ASDTests app, they can possibly be expanded upon further by implementing the current models that were determined with 100% accuracy. The potential mobile app could focus with an emphasis on toddlers to implement this highly accurate predictive model with the means of helping parent-child pairs to diagnose those before it's too late.

References

Thabtah, F. F. (2018, July 23). Autism screening data for Toddlers. Kaggle. https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers

Leroy, G. (2020). Using Natural Language Processing to Improve Autism Spectrum Disorder Research and Care | Digital Healthcare Research. https://digital.ahrq.gov/2020-year-review/research-summary/using-natural-language-processing-improve-autism-spectrum-disorder-research-and-care

Acknowledgments

I would like to thank my mentor, Dr. David Johanssen for providing me the necessary resources for this project and assistance with his very helpful input on the data. He helped a lot with pointing in the right direction this project would go and his supportive insight through each step along the way.

I also want to give thanks to Dr. Fadi Fayez Thabtah who helped provide the test data for this project and the various contributors at Kaggle who also shared and analyzed the data. They helped to provide plenty of tested observations for these datasets, which could be cross-referenced to verify their accuracy.