Autism Disparities In the World

Walla Rahama, Ankit Adimala, Shiwei Wang, Cam Powell, Rohan F., Dhruv Rokkam, Rahul D’Mello

Northeastern University, Boston, MA, USA

Abstract

This study aims to investigate disparities in Autism Diagnoses worldwide and identify the country with the highest proportion of cases, as well as use machine learning to attempt to predict suspected autism spectrum disorder according to the widely used AQ-10, perform regression analysis to compare ages of diagnoses, and compare cases and publications relating to across the globe. The project will involve utilizing pandas dataframes and DBeaver databases, along with regression analysis, random forest classifiers,, and other machine learning tools. Through these methods, we will generate general statistical analyses, create world map illustrations, and make predictions using machine learning algorithms.

Introduction

Autism spectrum disorder (ASD) is a complex neuron and psychological disorder. It included the symptoms of difficulties in social interaction, communication, and repetitive behaviors (Centers of Disease Control and Prevention, 2022). However, the diagnosis of autism is very different from region to region, mainly caused by the different diagnostic criteria, cultural influences, and public health services. Some previous research indicated that differences in diagnostic patterns by ethnicity suggested possible variations in parents’ descriptions of symptoms, clinician interpretations and expectations, or symptom presentation (Madell, 2006). Therefore, the goal of our project is to identify the disparities of autism diagnosis worldwide, focusing on the difference between each region. Analyzing global data could assist in identifying the factors contributing to these disparities. And by performing analysis on the AQ-10, we can also find ways the current diagnosis process may be impacting these disparities. Who are these tests designed or more suited for?   In this project, we are going to use pandas dataframe, DBeaver database, and machine learning tools to support our study.

Data Sources and Methods

AQ-10

The AQ-10 is a widely used short ten question survey, often administered by primary care practitioners to refer patients for ASD assessment. The ten questions asked in the survey vary depending on age in order to assimilate questions that address the most significant concerns for autistics in a particular age group. Here are the list of questions for both our child and adult datasets:

Children (4-11yo) AQ-10 :

1) S/he often notices small sounds when others do not.
2) S/he usually concentrates more on the whole picture, rather than the small details
3) In a social group, s/he can easily keep track of several different people's conversations.
4) S/he finds it easy to go back and forth between different activities
5) S/he doesn't know how to keep a conversation going with his/her peers
6) S/he is good at social chit-chat
7) When s/he is read a story, s/he finds it difficult to work out the character's intentions or feelings
8) When s/he was in preschool, s/he used to enjoy playing games involving pretending with other children
9) S/he finds it easy to work out what someone is thinking or feeling just by looking at their face
10) S/he finds it hard to make new friends

Scoring Criteria: Only 1 point can be scored for each question. Score 1 point for Definitely or Slightly Agree on each of items 1, 5, 7 and 10. Score 1 point for Definitely or Slightly Disagree on each of items 2, 3, 4, 6, 8 and 9. If the individual scores 6 or above, consider referring them for a specialist diagnostic assessment.

Adult (18+yo) AQ-10:

1) I often notice small sounds when others do not
2) I usually concentrate more on the whole picture, rather than the small details
3) I find it easy to do more than one thing at once
4) If there is an interruption, I can switch back to what I was doing very quickly

5) I find it easy to 'read between the lines' when someone is talking to me
6) I know how to tell if someone listening to me is getting bored
7) When I'm reading a story I find it difficult to work out the characters' intentions
8) I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant etc)
9) I find it easy to work out what someone is thinking or feeling just by looking at their face
10) I find it difficult to work out people's intentions.

Scoring Criteria: Only 1 point can be scored for each question. Score 1 point for Definitely or Slightly Agree on each of items 1, 7, 8, and 10. Score 1 point for Definitely or Slightly Disagree on each of items 2, 3, 4, 5, 6, and 9. If the individual scores 6 or above, consider referring them for a specialist diagnostic assessment.[3][4]

Data Collection

Dr. Fadi Thabtah is a researcher with a PhD in computing and mathematics as well as a masters in health psychology. Dr. Thabtah's main research interests lie in Big Data Analytics, Machine Learning and their implications in medical informatics. He has developed multiple screening applications for both Autism as well as Dementia and Alzhemiers in order to make diagnoses more accessible, and also collect information to help improve the current Autism diagnosis process. One of those apps is called ASDTests, a mobile app, that allows individuals, parents, and academic researchers to help recognize common autistic traits. The app uses the AQ-10, which stated previously, was designed to target "red flag' symptoms, and tends to ignore the many nuances in the autistic experience. These datasets also include demographic data, like gender, age, ethnicity, and place of residence. Both datasets for children and adults that were collected were products of ASDTests, which is important to keep in mind going forward. All the data received from ASDTests is reliant on mostly self-reported data, which may be exaggerated to a certain extent[5][7][8]. Additionally, a third dataset was retrieved from the U.S. Department of Health & Human Services. The department put together a dataset of information from peer-reviewed Autism prevalence studies by compiling publications based on the following criteria: published in English, produced at least one Autism prevalence estimate, and population-based with a defined geographic area. All of our data was imported into three different tables in a Sqlite database using DBeaver.

Methods:

Logistic Regression

We wanted to investigate the impact of different variables such as age, gender, and ethnicity when determining if a person has autism or not. In order to do this, we decided to use a logistic regression as this model is well-suited for predicting binary outcomes. In this instance, our dependent variable would be if the person has autism or not based on the questions, and our independent variables would be age, gender, and ethnicity. In order to understand which genders and ethnicities would have a greater impact on autism diagnoses, we decided to create dummy variables for each gender and ethnicity. This would allow us to get coefficients for each gender and ethnicity, allowing us to investigate the significance of gender and ethnicity. We also wanted to investigate how these variables would vary in significance when it came to adults and children, so we created two logistic regressions, one focusing on the adult dataset and the other focusing on the child dataset. From the results of the regression, we created two coefficients plots- one for adults and one for children - depicting the coefficients for each variable.

Machine Learning on AQ-10 Questions

One important feature in the dataset that we did not want to overlook were the survey questions, since the survey answers were then inferred, given a score to help determine whether or not there might be signs of autism. Since the scoring is solely dependent on the results of all ten questions, we decided to train our data based on the implications of question subgroups, agree or disagree. In both the child and adult AQ-10 questionnaire, four questions answered with a Slightly Agree or Strongly Agree are weighted towards our overall result score, while six questions answered a Slightly Disagree or Strongly Disagree. Our agree questions are more geared towards sensitivity and strengths found in autism, and our disagree questions are more focused on impairments and variance from the neurotypical experience. Training on these subgroups instead of the whole set of questions gives our result relevant meaning, for our ASD classification is not completely rooted in reality and solely operates on questionnaire responses. Instead, the AQ-10 should act as a precursor to encourage individuals to take more steps in achieving an official diagnosis. If our classifiers are able to successfully predict a result solely based on one subgroup of questions, we can make the implication that the subgroups of questions are somehow interconnected or dependent. Inferring that the increased neurological sensitivity and hyperactivity associated with autism goes hand in hand with the deviation of social, learning, and communication abilities between allistics and autists, rounding out a commonality in autistic experiences. Since we are working with two separate tables, for children and adults, each of the following methods should result in four different results or plots.

1. Random Forest Classifier/Feature Importance: We chose to use a random forest classifier to determine which questions in each subgroup of Dis/agree contribute to our resulting ASD classification. A classifier is needed here because we are targeting a binary attribute, ASD, using numerical data. I.e. our questions. By analyzing which questions hold more importance, we can compare what kind of questions seem most prevalent in both children and adults. In the future, this kind of analysis can aid in the continuous fight for improvements in early screening for Autism.
2. Decision Tree: We utilized a decision tree classifier to analyze the relationship between different questions and if that individual was suspected of having ASD or not by our AQ-10 questionnaire. Our random forest classifier may be more useful here since it is less robust to outliers, however; by using a decision tree as well we get more information on the classifier's process in the form of nodes. Ideally, based on how we grouped our data, each decision tree should be grouping the far right side of the network as majority no ASD while the far left should be classified as ASD. Our goal is to compare how our classifiers behave between the children and adult datasets as well as what subgroup of data makes for a better classifier.
3. Confusion Matrices: Lastly, our confusion matrices are used to measure the efficiency of the resulting decision trees. Here we can analyze and make educated inferences about factors that may or may not have impacted our classifiers' ability.
4. Aggregate Feature Importance: We created two types of feature importance plots for both of our datasets. The first was where we split the 10 questions by agree

and disagree and the second was all the questions aggregated together. Our second plot, which is a feature importance of all the questions, showed us very different conclusions than our first. By creating an aggregate feature importance we ignore the disagree or agree importance to the questions to see which question(based on the data) contributed to the recommendation.

Histogram & World Map

We are using the histogram to find the number of publication changes from 1966 to 2022. This will allow us to determine the increased speed and shifts in people's attention in this field. Additionally, the histogram helps us compare the research interests of different countries and the quantity of research, which supports our understanding of the disparities in research attitudes towards Autism.

Furthermore, we decided to use world-boundary data from geopandas to plot a map that depicts estimates of Autism prevalence per country from the publications dataset. After creating this map, we applied this same process to a combined adult and child dataset. Using their self-reported country of residence, we were able to plot the prevalence of Autism in different countries based on the proportion of people from a certain country that got diagnosed with ASD after taking the test. Based on the comparison of these two world maps, we can analyze the accessibility and diversity of the test by looking at over/underrepresentation.

Data Cleaning/Filtering:

Logistic Regressions

To carry out the logistic regressions, we had to use both the adult and child datasets. As the logistic regressions were only investigating variables such as age, gender, and ethnicity, we filtered the dataset by creating different views within the database so we could read the data from the view directly to a pandas dataframe. We created two views - one for the adult table, and one for the child table - which contained data on the patient type, age, gender, ethnicity, and whether or not the person has autism depending on their answers to the questions. Additionally, as many of the values in columns such as gender, ethnicity, and autism were of string data types, we decided to convert these values to a numeric data type in order to allow for the regression model to work as intended. For the gender, and ethnicity values, as explained before, we decided to create dummy variables, in order to convert them to a data type which would allow for the regression model to work, and to allow us to understand how different genders, and ethnicities contributed to overall autism diagnoses. We also converted ‘yes’/’no’ values within the autism column to 1/0 values to allow for the regression model to work.

Machine Learning on AQ-10 Questions

Because both our tables relating to ASDTests have little to no null values for the answers to the questionnaire implying the result score as well as the ASD classification also had little to no empty values, only light filtering was needed. In order to get rid of extreme outliers, rows were filtered so that result scores 0.0 and 10.0 were excluded from our data. Similarly to the logistical regression ‘yes/no’ values within the ASD column were changed to 1/0.

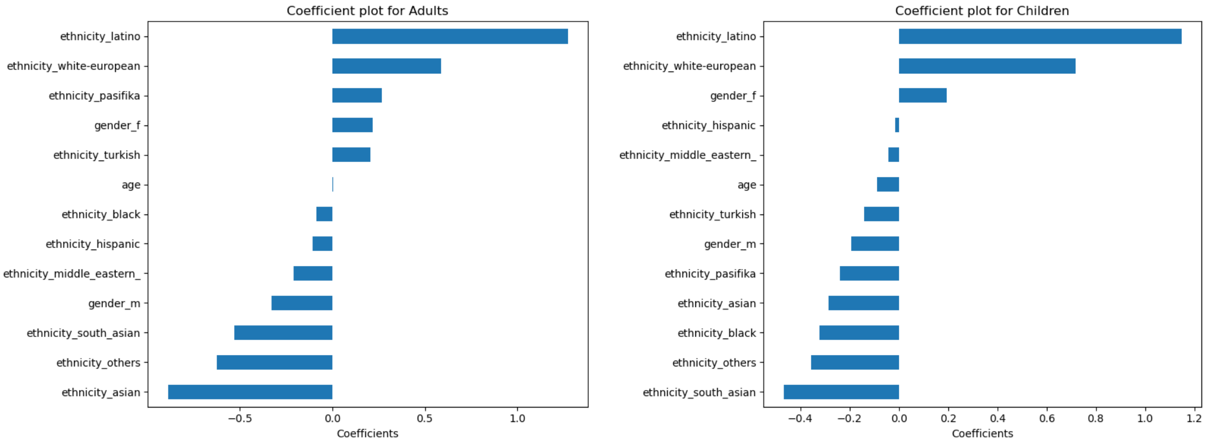
World Map on ASD Prevalence Estimates

Since the country specific data was most relevant to producing a world map, the dataset on publications was filtered for the different countries and their estimated ASD prevalence(individuals per 1,000). Since several countries had multiple publications, their estimates were averaged so that each country only appeared once in the filtered table. In order to plot these estimates on a map, the geopandas world library was required. This meant that the filtered table had to be cleaned so that the country names matched those in the geopandas library. For example, cleaning this table included renaming USA to ‘United States of America’, as well as separating the Caribbean into its individual countries, and combining the estimates for the countries that formed the United Kingdom. Other similar adjustments had to be made to the table. As for the second map, this one depicted ASD prevalence from the dataset on Adult and Child self reported results. To filter these tables, they were first combined and then a view was created with the relevant columns: country, number of ASD diagnosed individuals, and the total individuals from that country. These rows also had to be aggregated so that each country’s proportion could be viewed, which was done by summing the country’s totals into distinct rows. This table had to be cleaned in a similar manner to the table from the publications dataset; The countries listed were made to match those in the geopandas library. The proportion for each country was calculated by dividing the number of individuals diagnosed with ASD by the total number of people from that country. In order to get an estimate per 1,000, that proportion was multiplied by 1,000. However, a lot of countries were overrepresented and underrepresented, so the estimates were then divided by 40. This scaling was done so that the highest estimate per 1,000 was 25. This allowed the overestimation of ASD prevalence to be apparent while still being comparable to the world map from the publications dataset.

Analysis

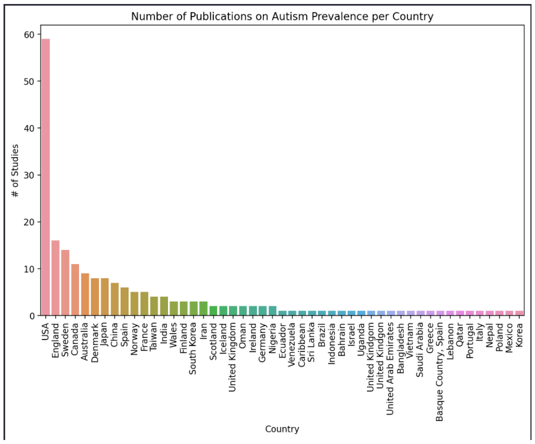
General Statistics

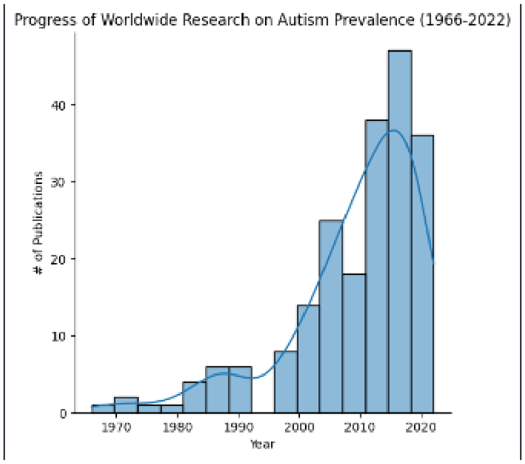
[Figure 0]



The graphs above depict the coefficient plots for the logistic regressions conducted on both the Adult and Child datasets. With the adult logistic regression, the model’s accuracy score was calculated to be 87%. This means that the model correctly predicted the outcome of the autism diagnosis 87% of the time. Similarly, the children's logistic regression model had an accuracy score of 84%. As can be seen from the coefficient plot for adults, several factors were seen to be statistically significant predictors of autism. These include certain ethnicities and genders, notably, being Latino, White-European, Pacifika, Turkish, or female were associated with higher odds of autism. Similarly, with the coefficient plot for children, being Latino, White-European, or female were associated with higher odds of autism.

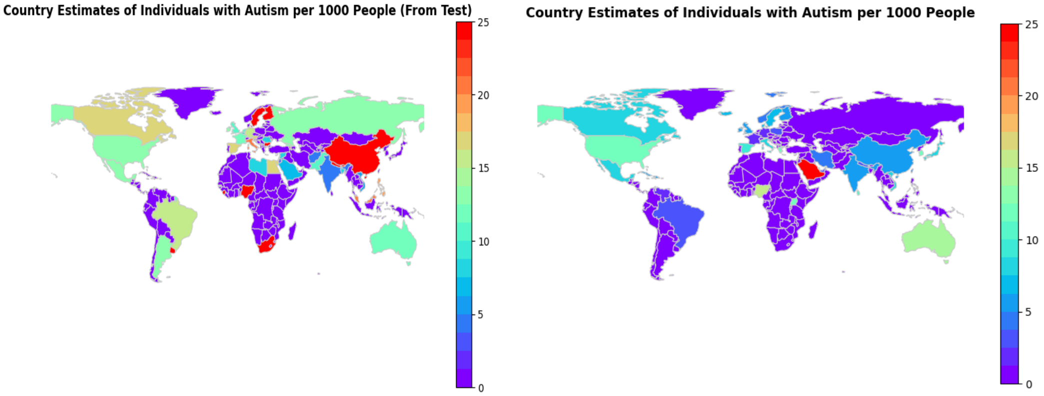
Histograms of Publications [Figure 1]





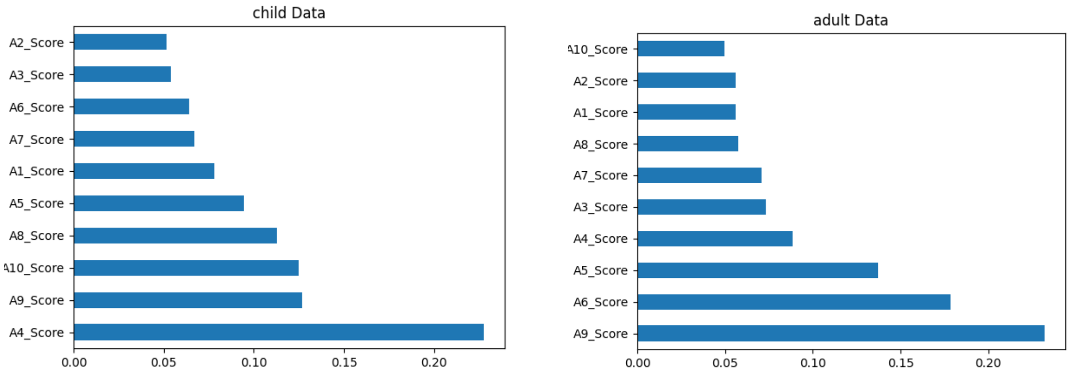
The histograms illustrated the number of publications per country and how overall research on Autism prevalence grew over the past 55 years.

World Maps [Figure 2]


Country Estimates of Individuals with Autism per 1000 People (From Test)
25
20
15
10
5
0
Country Estimates of Individuals with Autism per 1000 People
25
20
15
10
5
0

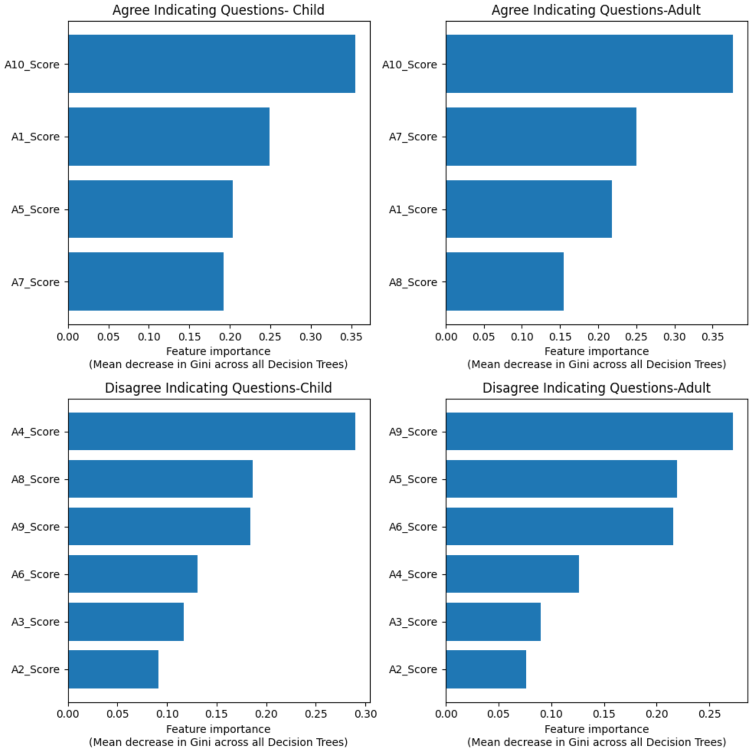
Graph 1 illustrates country estimates of individuals with Autism per 1000 people from the questionnaire-based data; Graph 2 illustrates country estimates of individuals with Autism per 1000 people from the research-based data. In Graph 1, the number of individuals with Autism are overestimated as compared to Graph 2.

Child and Adult Aggregate Feature Importance [Figure 3]



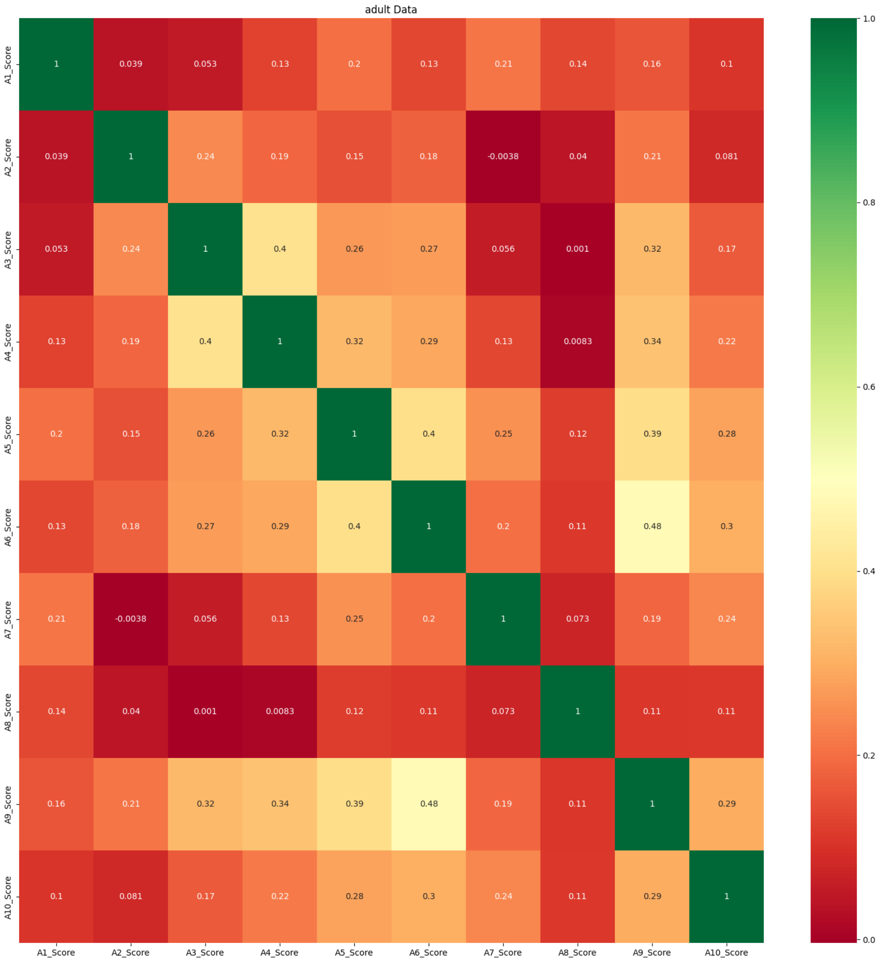
These visualizations display the importance of each question and how it correlates with the outcome of the AQ-10.

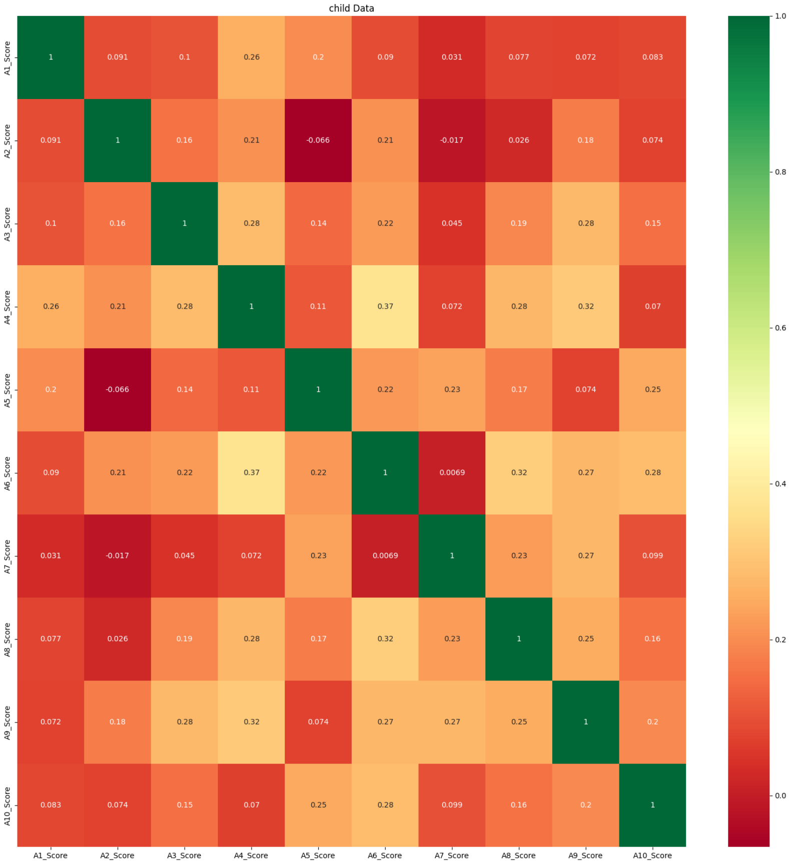
Feature Importances(Split by Agree and Disagree Questions) [Figure 4]



This visualization shows feature importance for questions split by disagree and agree importance, and also by the children and adult datasets. As seen in the charts, which questions were more influential depended on which dataset was being examined.

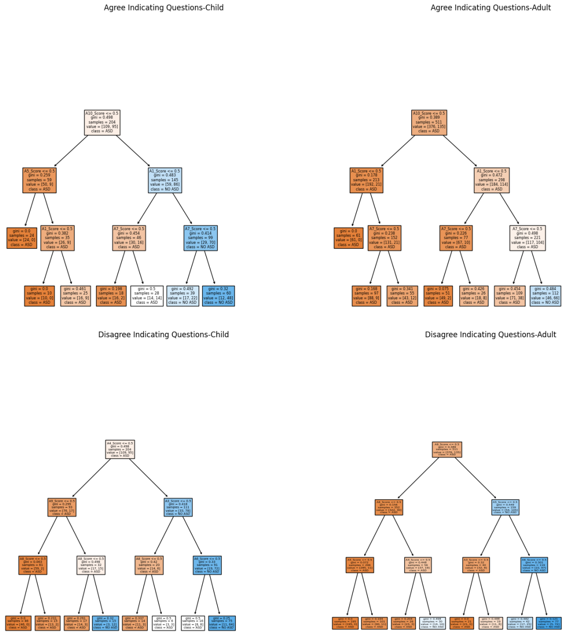
Question Confusion Matrix[Figure 5]





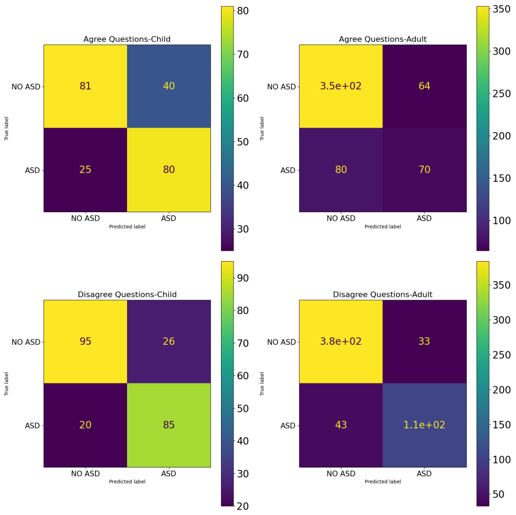
Above are two simple correlation plots of all 10 questions in the survey. While it doesn't reveal any casual evidence it does support our confidence in the validity of the data since some questions like A6(I know how to tell if someone listening to me is getting bored) and A9(I find it easy to work out what someone is thinking or feeling just by looking at their face), in the adult dataset, show strong correlations(.48).

Decision Trees [Figure 5]


Agree Indicating Questions-Child
Agree Indicating Questions-Adult
Disagree Indicating Questions-Child
Disagree Indicating Questions-Adult

The decision trees above are broken down into the two datasets we worked with, children and adults, and are also broken down into Agree and Disagree questions for four trees. These allow a viewer to see the paths between which answers were given and what decision resulted. All our decision trees had accuracies above 70%, with the most accurate being disagree questions for adults. Keep in mind, this subgroup has disproportionate 'NO ASD' classification contributing to its inflated accuracy.

Confusion Matrices [Figure 6]



In the above confusion matrices, we examined the efficiency of our decision tree classifier between questions, for which both Agree and Disagree question types show a stronger performance to the test taker receiving a recommendation for an autism diagnosis for the children dataset rather than for the adults. It is also worth noting the disagree questions made better classifiers than the agreed questions. One may make the inference that it may be because this subgroup consists of more questions. There is still adequate accuracy for the adult dataset, although not as strong as for the children dataset. It is also worth noting that the children dataset has a significantly smaller sample size than the adult dataset.

Conclusions

According to our data analysis, a significant increase in the publication of studies on Autism has been observed from 1966 to 2022, particularly since 2000. The increase in publications indicates a shift in attention on this disorder within the global community. The database revealed that the United States has the most research output, and other countries like England, Sweden, Canada, Australia have also shown considerable effort in the research field.

According to our illustrations, two world maps that depict the number of Autism cases illustrate that the United States and other European countries had relatively fewer cases compared to China, and Saudi Arabia, the two major countries with the higher number of Autism diagnosis. This finding suggests that there is no strong relationship between the number of Autism diagnoses and the number of research publications, while the level of research attention is not the major factor to influence the number of Autism diagnoses.

Additionally, the first map is the country estimates of Autism per 1000 people for questionnaire-based data, and the second map is the country estimates of Autism per 1000 people for research data [Figure 2]. Upon examining these two maps, the research-based map exhibits lower numbers of diagnoses across all countries compared to the questionnaire-based map. This finding suggests that the questionnaire-based world map may have overestimated the prevalence of Autism worldwide. The reason for this overestimation could be attributed to the limited accessibility of the questionnaire-based test to a diverse enough sample size.

Through our multiple feature importance plots, we can draw multiple clear conclusions. First, for both adults and children, question 10 had the highest mean decrease in gini score out of all the Agree questions. For adults, question 10 is: “I find it difficult to work out people’s intentions," and for children, question 10 is: "S/he finds it hard to make new friends." However, for disagree questions, adults and children begin to differ. For adults, question 9, which is "I find it easy to work out what someone is thinking or feeling just by looking at their face," had the highest mean decrease, while for children, it was question 4, which is "S/he finds it easy to go back and forth between different activities." Another interesting phenomena we found is that for both the agree and disagree questions, the feature importance between children and adults had very similar shapes. This should encourage a more content based analysis on the questions proposed on both versions of the AQ-10. For example, questions 8 and 9 on the children's questionnaire compared to 5 and 6 on the adult's AQ-10 share a theme of invisible social cues and plot similarly on our agree-disagree sectioned feature importance chart.

According to the coefficient plots, it is shown that the test works disproportionately better for white Europeans, latinos, and women compared to many of the other ethnicities tested for both the children and the adult datasets. This could be due to the sampling size of the data, such as simply having more white Europeans taking the test, and it is not possible to give a concrete takeaway. However, the results are interesting and reveal that it may be necessary to create a revised test that tests each ethnicity proportionally.

According to the decision tree visualizations, it becomes clear that the AQ-10 was much more effective for children than for adults. We would have expected more blue boxes on the bottom right of each tree, and more orange boxes on the bottom left of each tree. This is not the case for each tree, and is more true of the children than of the adults. It is also notable that the gini scores of many of the boxes is relatively high, which is not necessarily a good sign for the AQ-10. This is because if we were to add more depth to our trees, especially for the agreed questions, our classifiers would be overfit. This furthers our findings that the AQ-10 is more effective for use with children than with adults.

According to the confusion matrices, we can see that the children had a stronger performance on the test than the adults comparatively. The child classifiers were able to predict a larger proportion of the dataset with suspected autism correctly. Since ASD is a very nuanced disorder, it is more beneficial to push patients for further consultation even if they do not reach official diagnosis than lack of endorsement for those who may have ASD, but scored less than a six on our five minute questionnaire. We found this interesting as we would have expected adults to be more in touch with their behaviors and signs as they have gone through life with them compared to children. One possible explanation is that the adults have become accustomed to masking their autism and find it harder to self-recognize, however it is not possible to concretely state a cause.

Author Contributions (each write 1-2 sentences for our individual contributions)

Walla (Point Person) : Proposed research idea, including the responsibility of collecting our main datasets, as well as doing preliminary research on data collection, reliability, and general knowledge of ASD and its history, In regards to code, I produced all three visualizations under *Machine Learning on AQ-10 Questions.* In the report, I wrote about the AQ-10, Dr. Fadi Thabtah and his goals for ASDTests, the methods and data cleaning performed in regards to *Machine Learning on AQ-10 Questions,* as well as revisions through all sections. Led Day 2 of presentations. Generated backup SQL database.

Ankit: Created the relational database with the three necessary tables; Worked on data cleaning on the publication dataset; Wrote the code for histogram of publications, and heatmaps of the world ASD estimates. Worked on the slides and added the visualizations; Presented on the first day. Contributed to the method, data cleaning sections in the report.

Shiwei: Created the google doc, slides; Finished the writing on the abstract, introduction, analysis, and conclusion parts. Specifically writing the report focusing on the world maps and publications number. Presented on the second presentation day.

Cam: Created the presentation slide and compiled the visualizations used. Led the presentation on presentation day 1, introducing the project, our overall goals and data used, and explaining the confusion matrices on the slide. Contributed to the Abstract, Introduction, Analysis, and Conclusion sections of the report. Researched ASD and general knowledge of ASD. Aided in tweaking the output of visualizations used in the project.

Rohan: Created aggregated feature importance charts and correlation map. Presented said plots on the day of presentation and contributed to the methods and analysis part of the report.

Dhruv: Worked on data collection along with shifting between the different visualization groups and code-checking on deep note websites by adding comments and making sure the most efficient methods possible were used.

Rahul: Created the logistic regression statistics investigating the impact of variables such as age, gender, and ethnicity on determining autism diagnoses. Produced coefficient plots from the resulting statistics for both adults and children. Presented the coefficient plots on presentation day 1. Contributed to the methods, and analysis sections of the report.

References

1. Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Journal of the American Academy of Child and Adolescent Psychiatry, 51(2), 202-212.
2. Centers for Disease Control and Prevention. (2022, March 28). Signs and symptoms of autism spectrum disorders. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/ncbddd/autism/signs.html#:~:text=Autism%20spectrum%20disorder%20(ASD)%20is,or%20repetitive%20behaviors%20or%20interests.
3. Engelbrecht, N. (2022, August 20). The AQ-10. Embrace Autism. Retrieved from https://embrace-autism.com/aq-10/#test.
4. Mandell, D. S., Ittenbach, R. F., Levy, S. E., & Pinto-Martin, J. A. (2006). Disparities in diagnoses received prior to a diagnosis of autism spectrum disorder. Journal of Autism and Developmental Disorders, 37(9), 1795-1802. https://doi.org/10.1007/s10803-006-0314-8
5. Thabtah, F. (2023). Dr. Fadi Thabtah. Manukau Institute of Technology. Retrieved June 21, 2023, from www.manukau.ac.nz/study/areas-of-study/digital-technologies/meet-the-team/digital-technologies-team/fadi-fayez
6. Thabtah, F. (2017). Autistic Spectrum Disorder Screening Data for Children. UCI Machine Learning Repository. https://doi.org/10.24432/C5659W.
7. Thabtah, F. (2017). Autism Screening Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5F019.
8. U.S. Department of Health & Human Services. (2020, November 10). Autism Prevalence Studies. Catalog. Retrieved from catalog.data.gov/dataset/autism-prevalence-studies