Diagnosing Autism in Demographics

Patrick Shan
patricklshan@gmail.com
(650) 334-9662
September 3, 2019

I. Introduction

Problem

Given the rising number of global Autistic Spectrum Disorder (ASD) cases, there is interest in developing new methods to diagnose this condition. The Manukau Institute of Technology in Auckland, New Zealand conducted a worldwide survey for ASD diagnosis. Although the survey offers some preliminary insight into autism prevalence, there are some unanswered questions such as whether certain ethnicities or gender are more likely to show symptoms of ASD. This project seeks to study the ethnic and gender factors that affect the likelihood of showing symptoms or becoming diagnosed with ASD.

Approach

The approach for this project would be comparing the rates of autism diagnosis and questionnaire scoring for symptoms in each ethnic and gender group. The approach itself would be performed using data wrangling, exploratory data analysis (EDA), and machine learning. In data wrangling, the dataset would be organized into a table that is indexed and cleaned for consistency. After data wrangling, EDA would visualize potential data trends and perform statistical analysis to test these observations. Finally, machine learning would validate the outcomes of EDA and determine the weight of different demographics features. This approach would determine which demographic features are most likely to be associated with autism.

Clients and Applications

The main clients of this study would be social workers and autism researchers. By identifying demographics who are most likely to be diagnosed with ASD, social workers can concentrate their attention on these communities, thereby improving the effectiveness of their efforts. Researchers would also benefit from the results as they can identify the demographics most likely to be associated with autism and conduct more research on these demographics of interest. Furthermore, if any defects and sampling errors are found in the original dataset, the original owners would benefit from improvement advice for their future surveys. The results of this study will help improve the efficiency and resource allocation for the clients.

II. Data Collection and Wrangling

Data Sources

The dataset was originally posted on Kaggle by the Manukau Institute of Technology from Auckland, New Zealand. The data itself is a questionnaire conducted on over 700 individuals around the world, and it asks for ASD diagnosis status along with 20 separate attributes. Given the relative large size of the data sample, errors would have less of an impact and this would allow for different machine learning models that wouldn't be possible with a smaller dataset. Furthermore, since this dataset contains demographic information, no additional datasets were needed as references.

Data Cleaning Compilation

Before conducting data analysis, the dataset was cleaned and reorganized in Jupyter notebooks using Python as the coding language. In Jupyter, the dataset to converted to a pandas dataframe format to allow for clean editing. The columns were renamed to remove unnecessary punctuation, correct spelling, and merge columns. Furthermore, the ethnicity names were standardized to have consistent spelling, punctuation, and capitalization, thereby reducing redundant values. The features of interest for this study are ethnicity, gender, ASD diagnosis status, and scoring. Two smaller dataframes were created from the main dataframe and are grouped by either ethnicity or gender. Each of these smaller dataframes contains information about the percentage of people with autism, the total number of people with autism, and the mean questionnaire scores. From these three main dataframes, the visualization and statistical analysis will examine sample proportion with autism (rates of ASD diagnosis) and scoring (higher values are associated with more noticeable symptoms).

III. Exploratory Data Analysis

Hypothesis

The hypothesis for exploratory data analysis (EDA) is that individuals of male gender and white-european ethnicity have higher scores and ASD diagnosis rates. Autism is generally assumed to be genetically inherited and men are historically more likely to inherit genetic disorders. Furthermore, white caucasians are more likely to be diagnosed with autism since this condition has a greater history of documentation in this ethnic group. As such, this would be the likely pattern discovered in the research.

Visualization

Visualization through graphing provides a visual aid for detecting patterns and allowing for comparison between demographics. Bar graph was the graph type used in this analysis since it would show the quantitative differences in values. The values for gender and ethnicity were graphed separately from each other to avoid interference and used the smaller grouped-by datasets created during data wrangling. By graphing the scores and diagnosis rates for genders and ethnicities, the visualization would show if there are any tangible differences among demographics.

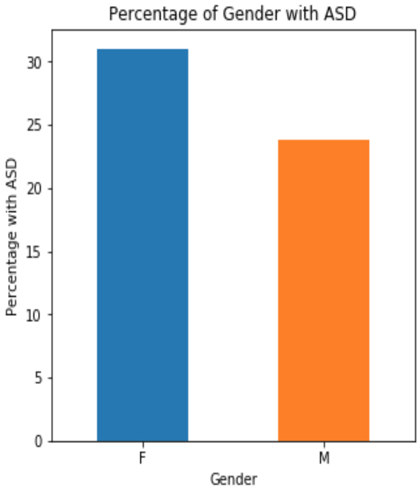

Figure 1. Percentage Composition of Gender with ASD. Women (represented as F) have higher rates of ASD-diagnosis than men (represented as M).

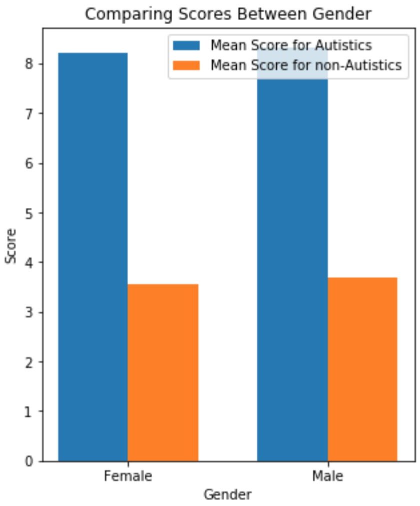

Figure 2. Comparable Scores Between Gender. There is no difference in score between the genders. Autistics have higher scores than non-autistics.

The visualization for gender offers some interesting insights into the differences between the two genders when it comes to diagnosis rates and scoring. The visualization for gender proportions showed that women have higher diagnosis rates than men. Furthermore, the scores of individuals with ASD are higher than those of individuals without ASD. However, there is also no difference between the genders in regards to scoring. The visualization implies that while women are more likely to be diagnosed with ASD, they aren't more or less likely to show ASD symptoms than men.

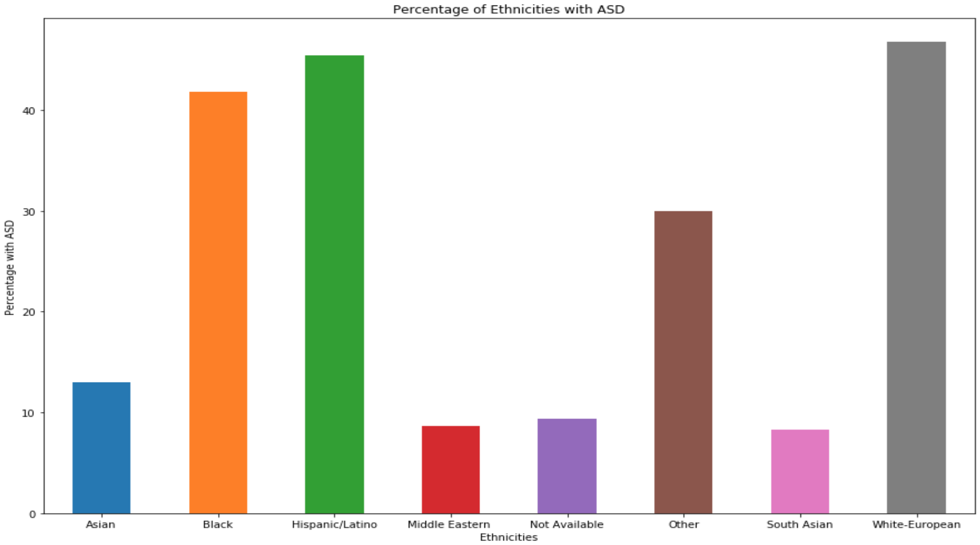

Figure 3. Percentage Composition of Ethnic Groups with ASD. White-European, Hispanic/Latino and Black are the ethnicities have the highest population proportions with ASD. In contrast, Middle Eastern and South Asian have the lower population proportions with ASD.

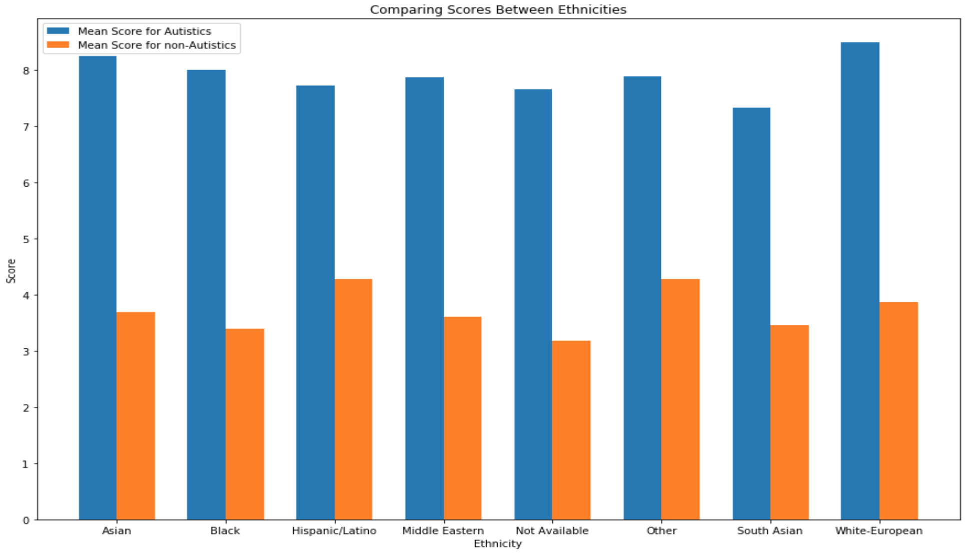

Figure 4. Comparable Scores Between Ethnicities. The scores among autistics is relatively consistent between ethnicities. Furthermore, the scores among non-autistics is also relatively consistent between ethnicities. On average, the scores of autistics are roughly double that of non-autistics.

The visualization of ethnicity offers a more complicated picture given the presence of so many different ethnic groups with varying sizes. While White-European is the most represented ethnicity in the overall dataset, White-European, Hispanic/Latino and Black are the ethnicities have the highest population proportions with ASD. As with the gender comparison, the proportion of individuals with ASD is nearly double that of individuals without ASD. While White-European, Hispanic/Latino and Black are the ethnic groups have the highest mean scores, the mean scores of the ethnicities show relatively little variation between ethnicities. Furthermore, the scores among all non-ASD ethnicities are relatively consistent. These trends seem to suggest that White-European, Hispanic/Latino and Black are the most likely to have ASD symptoms and diagnosis rates.

Descriptive Statistics

Descriptive statistics expand upon the patterns that were seen in the visualization. For gender populations, females have higher proportions than males. However the mean scores among individuals with autism do not differ according to gender. Among ethnic groups, Black, Hispanic/Latino, and White-Europeans have the highest rates of ASD diagnosis and mean score. These results identify groups of interest and potential trends that could be further explored.

Table 1. Descriptive Gender Statistics. This table shows many attributes of both gender-based sample groups. The main parameters are percentage with autism, mean score for people with autism, and mean score for people without autism.

Gender	Percentage with Autism	Mean Score for People with Autism	Mean Score for People without Autism
Female	31	8.21	3.57
Male	23.81	8.31	3.71

Table 2. Descriptive Ethnicity Statistics. This table shows many attributes of both ethnic-based sample groups. The main parameters are percentage with autism, mean score for people with autism, and mean score for people without autism.

Ethnicity	Percentage with Autism	Mean Score for People with Autism	Mean Score for People without Autism
Asian	13.01	8.25	3.68
Black	41.86	8	3.4
Hispanic/Latino	45.45	7.73	4.28
Middle Eastern	8.70	7.88	3.61
Not Available	9.38	7.67	3.17
Other	30	7.89	4.29
South Asian	8.33	7.33	3.45
White-European	46.78	8.50	3.87

Inferential Statistics

However, while descriptive statistics could identify trends and patterns, inferential statistics are needed to verify the validity of these trends. The statistical test chosen is the two-sample z-test, since the sample variance can be deduced and the size of the autism data is greater than 30 samples. The z-test is used for comparing two separate datasets and determine whether the datasets are statistically significant based on the calculated p-value. If the p-value is less if the p-value in a test is less than 0.05, then it means that the two datasets are statistically significantly different. This process is used to compare sample proportion with autism and autism score within the gender and ethnicity sub-dataframes.

Table 3. Inferential Gender Statistics. This table shows the p-values for comparing proportion and score among autistics between genders. A p-value less than 0.05 indicates that there is a statistical difference between genders.

P-Value for Proportion	P-Value for Score Among Autistics
0.035	0.519

Table 4. Inferential Ethnicity Statistics. This table shows the p-values for comparing proportion and score among autistics between each ethnic group and the general population. A p-value less than 0.05 indicates that there is a statistical difference between the ethnic group and the general population.

	P-Value for Proportion	P-Value for Score Among Autistics
Asian	0.001	0.996
Black	0.039	0.286
Hispanic/Latino	0.023	0.019
Middle Eastern	1E-4	0.251
Not Available	1E-4	0.009
Other	0.741	0.214
South Asian	0.012	0.001
White-European	3.595E-8	0.017

The inferential statistical analysis of gender was relatively straightforward when it comes to producing results. When it comes to comparing sample proportion with autism between the genders, the p-value is less than 0.05, indicating that such difference is significant. However, the testing between genders for score among autistics yields a p-value greater than 0.05, indicating that there is no significant difference. Therefore, the test shows that there is a difference between gender for likelihood of autism but that gender disparity is not present in the differences in score of autistics. Thus, this backs the inferential statistic assumption that women are more likely to have autism but there is no gender difference when it comes to displaying symptoms.

However, inferential statistical analysis of ethnicity was more complicated. Given how there were 10 separate ethnic groups, there wasn't a clear control for direct comparisons. To remedy this situation, multiple z-tests were conducted by pitting each of the ethnicities against either general population for diagnosis rates or among autistics for scoring. In testing for proportions with autism, all ethnicities barring Other have p-values less than 0.05. Furthermore, in testing for scoring among people with autism, Hispanic/Latino, Not Available, South-Asian, and White-European are the ethnicities whose p-values are less than 0.05. These studies seem to imply that Other are less likely to have autism but also autistics in the Hispanic/Latino, Not Available, South-Asian, and White-European ethnicities are more likely to show more symptoms.

Exploratory Data Analysis Summary

The visualization and statistical analysis of the EDA offer some interesting insights into diagnosis rates and scoring among the demographics. The visualization showed that for both gender and ethnicity compassions, individuals with ASD have higher scores than individuals without ASD. Visualization shows that Black, Hispanic/Latino, and White-European ethnicities have the highest diagnosis rates yet any differences in scoring between ethnicities do not seem apparent. Furthermore, the visualization shows that women have higher diagnosis rates than men yet there is no difference in scoring between the genders. As such, the visualization shows that there are differences between the various genders and ethnicities when it comes to ASD diagnosis rates yet there doesn't seem to be any differences in scoring.

Both the descriptive and inferential statistics offered some additional insights into the patterns as seen in the visualization. Statistical analysis confirmed that the differences in ASD diagnosis rates for gender is significant while also confirming that there is no significant differences between the genders when it comes to scoring. When compared to the mean ASD diagnosis rate of the entire sample, all of the ethnicities barring Other are significantly different from the mean value, thereby validating any differences between the ethnicities in this parameter. However, in testing for scoring among people with ASD, Hispanic/Latino, Not Available, South-Asian, and White-European are the ethnicities are more likely to have scores differing from the mean. This would imply that not only do ASD rates vary by ethnicity, but that also that certain ethnicities are more likely to show symptoms of ASD.

IV. Machine Learning

Hypothesis

The machine learning should support the observations inferred from exploratory data analysis (EDA). As such the models should show that White-European are more likely to have higher questionnaire scores and greater proportions with ASD than other ethnicities. Furthermore, the models should show that women have higher proportions but similar scores compared to men.

General Approach

The machine learning for the Capstone project is comprised of regression analysis and Naive Bayes classification. The regression analysis itself consists of both multivariate and forest models that will identify which demographic factor most closely correlates with the questionnaire score. Naive Bayes classification will determine the probability of ASD diagnosis depending on demographics. For all models, the independent and dependent variables are each split into training and testing portions with the training portions getting fitted to either a regressor or classifier; by splitting the data and fitting them to the right model, the machine learning is able to yield more accurate results. All models of machine learning were performed using ethnicity and gender variables.

Regression Methods

Linear Regression identifies the correlation between demographic factors and scoring. As the scores are continuous numerical values, linear regression analysis was chosen since it examines correlation between variables and numerical values. However, since the independent variables of demographics are categorical, they were first converted to dummy variables. After the conversion, the linear regression tests whether the independent variables of gender and ethnicity show correlation with the ASD score. A subsequent analysis computes the R^2 score and Residual Mean Square Error (RMSE); an R^2 value closer to 1 would indicate closer fitting regression while a low RMSE score indicates minimal deviation and predicted data reliably adhering to the fitted regression pattern.

Random forest regression is a supplementary model that finds the importance of demographics on scoring. Random forest itself is an aggregation of multiple decision trees, each of which shows branching paths to possible outcomes based on certain criteria (nodes). The main advantage provided by random forests is that it reserves data for observations and error estimation, ensuring more accuracy and less overfitting. For this project, the data is split into training and testing portions that are fitted to a regressor, which in turn can predict the weight of the demographic features in the data. Based on the results of the outcomes, the random forest regression can be used to the probability of these outcomes occurring and the importance of the different demographic features.

Table 5. Linear Regression Coefficients. This table shows the coefficients of the different demographic variables. Higher coefficients indicate strong, positive correlation between scoring and variable.

Demographic Variable	Coefficient
Asian	0.05
Black	-0.06
Hispanic/Latino	-0.06
Middle Eastern	-0.12
Not Available	-0.28
Other	0.37
South Asian	-0.26
White-European	0.36
Female	-0.04
Male	0.04
Austistic	2.24
Non-Autistic	-2.24

Table 6. Random Forest Regression Importance. This table shows the importance of each demographic variable on scoring according to random forest regression.

Demographic Variable	Importance
Asian	0.01
Black	0.11
Hispanic/Latino	0.12
Middle Eastern	0.01
Not Available	0.03
Other	0.05
South Asian	0.01
White-European	0.57
Female	0.04
Male	0.04

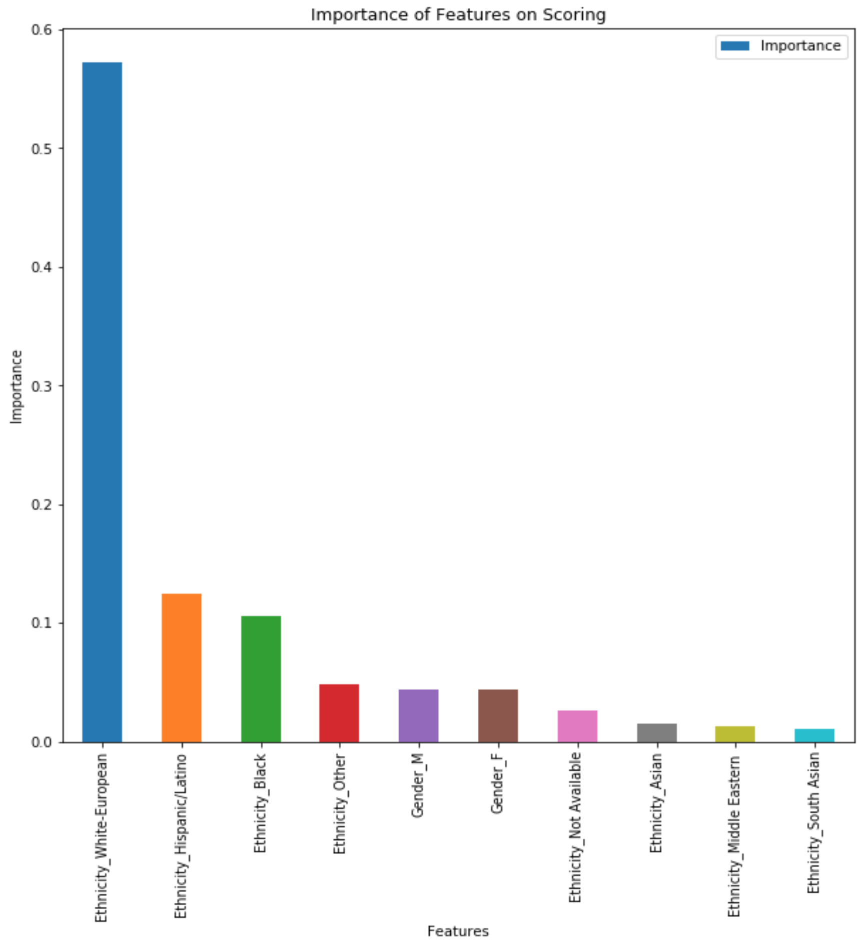

Figure 5. Importance of demographics on scoring. This graph shows the importance of difference demographics on scoring according to the random forest regression. White-European has the highest impact on scoring followed by Hispanic/Latino and Black. Both genders are the 4th and 5th most important demographics.

Regression Results

The linear regression model shows the demographic parameters most likely to affect individual scoring. The model has an R^2 value of 0.70 and RMSE of 1.36, suggesting that it is fairly reliable. ASD-positive individuals have the highest coefficient values at 2.23. Among ethnicities, White European, Hispanic/Latino, and Other have the highest respective coefficient values of 0.34, 0.10, and 0.32. This implies that White European and Hispanic/Latino are the ethnicities most likely to correlate with higher scores.

The random forest regression tests the importance of demographics on overall scoring. The model showed that White European ethnicity has the greatest impact with an importance factor of 0.57. However, while the test ran successfully in Python, the actual decision trees couldn't be successfully visualized due to issues with coding language compatibility. As such, this test seems to factor out ethnicity, specifically that of White European, as having the greatest impact on scoring.

Classification Methods

Naive Bayes (NB) is a classifier that determines the likelihood of an event occurring. For this project, the NB will determine the probability that an autstic individual belongs to a specific demographic and the probability that an individual in an ethnic group will be diagnosed with autism. Multinomial NB is used for ethnicity (allows multiple variables) while Bernoulli NB is used for gender (works best for binary variables). Both models will offer insights into the likelihood of autism diagnosis.

Random forest classifier was used for determining the importance of demographic features on ASD diagnosis. Like the random forest regression, random forest classifier relies on an aggregation of decision trees that show branching paths to possible outcomes. However, whereas the random forest regression was used for associating demographics with score, which in turn correlates with ASD symptoms, the random forest classification was used for associating demographics with ASD diagnosis.

Table 7. Ethnic Composition for Multinomial Naive Bayes. This table shows the ethnicity composition of ASD sample population and ASD rates in each ethnic group.

Ethnicity	Percentage of Autistics Comprised of Ethnic Group	Percentage of Ethnic Group with Autism
Asian	18.61	28.29
Black	8.14	35.40
Hispanic/Latino	6.98	39.54
Middle Eastern	9.30	18.91
Not Available	13.95	27.18
Other	4.65	28.99
South Asian	4.65	24.16
White-European	33.72	27.06

Table 8. Gender Composition for Bernoulli Naive Bayes. This table shows the gender composition of ASD sample population and ASD rates in each gender.

Gender	Percentage of Austistic Comprised of Gender	Percentage of Gender Diagnosed with Autism
Female	50.581	28.750
Male	49.419	25.886

Table 9. Random Forest Classification Importance. This table shows the importance of each demographic variable on scoring according to random forest classification.

Demographic Variable	Importance
Asian	0.02
Black	0.13
Hispanic/Latino	0.18
Middle Eastern	0.03
Not Available	0.03
Other	0.04
South Asian	0.01
White-European	0.46
Female	0.05
Male	0.05

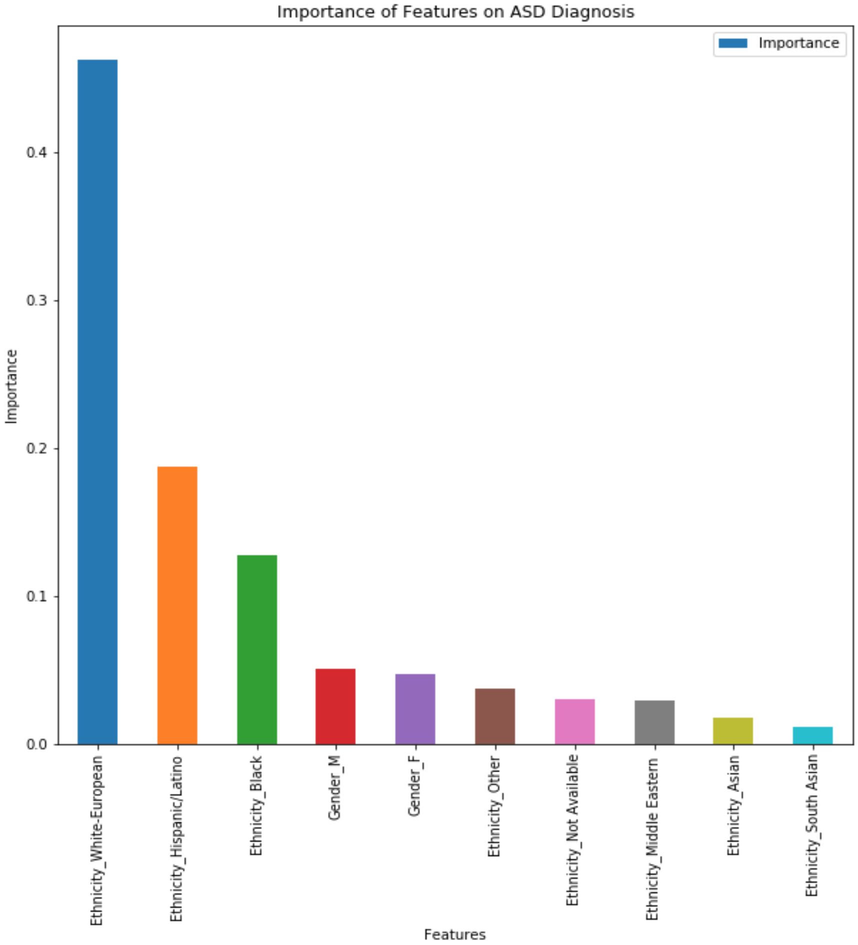

Figure 6. Importance of demographics on diagnosis. This graph shows the importance of different demographics on diagnosis according to the random forest classification. White-European has the highest impact on scoring followed by Hispanic/Latino and Black. Both genders are the 3rd and 4th most important demographics.

Classification Results

The Naive Bayes (NB) classifier examines the demographic composition of autism research group and the probability of demographic having autism. The multinomial NB on ethnicity shows that among the sample population diagnosed with ASD, Asians and White-Europeans make up the largest proportion at 18.61% and 33.72%, respectively. Conversely, Hispanic/Latino and Black are the ethnicities most likely to be diagnosed with autism at 39.54% and 35.4%, respectively. This indicates that while certain ethnicities make up larger portions of the ASD-positive sample group, others have higher rates of ASD diagnosis.

The Bernoulli NB on gender offers similar insights into autism composition albeit on a more reliable scale. The Bernoulli NB showed that among individuals diagnosed with ASD, 50.58% are women and 49.42% are men. Furthermore, the Bernoulli NB also showed that 28.75% of women and 25.89% of men are diagnosed with ASD. Furthermore, while the gender While these results seem to support the EDA patterns, the difference between gender seem less pronounced.

The random forest classification tests the likelihood of various demographics affecting ASD diagnosis. The test showed the importance of different demographics with White European ethnicity having the greatest impact with an importance factor of 0.44. While the test ran successfully in Python, the actual decision trees couldn't be successfully visualized. As such, this test seems to suggest that ethnicity, specifically that of White European, as having the greatest impact on successful diagnosis.

Machine Learning Summary

The Machine Learning analysis show some interesting trends and patterns in the autism data that could be contrasted with the EDA. The linear regression and random forest regression analyses show that ethnicity plays a greater role than gender with White European having the most importance followed by Hispanic/Latino and Black ethnicities. The Naive Bayes show that White-Europeans are the most represented ethnicity among autistics, the percentage of autistics among all ethnicities are less varied compared to the EDA results. Furthermore, while the Naive Bayes show that women are more likely than men to be diagnosed with ASD, the disparity between gender is less pronounced than implied by the EDA. As with the regression analysis, the random forest classification shows that ethnicity plays a greater role than gender with White European having the most importance followed by Hispanic/Latino and Black ethnicities.

The machine learning confirms and disproves several aspects of the hypothesis. The models confirm that women have higher ASD rates and that there is no difference between the genders in regards to scoring. The models also confirm that White-Europeans are the most likely to have higher scores and hold the most importance in both scoring and diagnosis rates. However, the results also disprove the hypothesis that White-Europeans have higher ASD rates as the disparities do not appear as drastic and some ethnicities have comparable ASD rates.

V. Conclusion

Summary

The goal of this study is to identify the demographics that are most likely to be associated with ASD diagnosis and symptoms. The data visualization offered some insights into gender and ethnicity demographic features. Graphs show that while women have higher rates of ASD diagnosis but there is no difference in scoring between the genders. The visualization showed that White-Europeans, Hispanic/Latino, and Black have the highest rates and scores among the ethnicities even though the number of individuals identifying as Hispanic/Latino and Black is smaller than White-Europeans. Statistical analysis support many of these observations; not only did the ethnicities have significant differences in rates and scoring, but the difference in gender rates are also significant. Both the visualization and statistical analysis of the EDA confirm that there are differences between the demographic features.

Machine Learning offered further insights into the patterns identified in EDA. Naive Bayes classification support many of the scoring and rates as identified in the EDA, though the disparity between ethnicities and gender seem less pronounced. Random Forest Regression and Classification showed that White-European ethnicity is the demographic with the most importance in regards to diagnosis rates and scoring; other demographics of importance include Hispanic/Latino and Black. The Random Forests also showed that gender has only a moderate importance. Subsequently, while the machine learning supports demographic patterns of EDA, it also shows that there were possible biases regarding the demographics.

Based on the results and analysis, there are some patterns regarding demographics and ASD that could be extrapolated. Ethnicity plays a greater role than gender in ASD diagnosis rates and symptom scoring. White-Europeans, Black, and Hispanic/Latino are the ethnicities with the most weight. Females have higher diagnosis rates than men yet there is no difference between gender in regards to scoring.

Limitations and Future Directions

However, there are future directions and discrepancies that must be taken into account. The main issue with the original dataset is how it suffers from biased demographics. Although the data was intended as a comprehensive overview of autism in the general global population, certain demographics are larger than others. This overrepresentation can lead to disparities between EDA and machine learning and make it more difficult for direct comparisons between demographics.

Many aspects like percentage composition of ethnicities seem to differ between the EDA and machine learning. The actual algorithms seem reliable given how they have training and testing scores of approximately 0.7. This may come down to data collection and the composition of the original data. The original dataset shows inconsistent sample size of the different demographics that can lead to inconsistent results. There are some solutions for future studies such as ensuring equal population sizes across ethnicities or consolidating individual ethnic groups into larger regional groups. By ensuring that the sample sizes are larger and more consistent, the analysis could be more reliable.

Furthermore, the Random Forest Regression and Classification show that there may be potential biases in regards to certain ethnicities. While White-European ethnicity is shown to have the most impact on rates and scoring, much of this comes to higher overall representation. In the original data set, 233 out of the 686 total participants belong to the White-European ethnicity. Furthermore, the visualization and statistical analysis showed that Hispanic/Latino and Black have comparable rates of ASD diagnosis despite having smaller representation. Given how the data is intended to be a comprehensive overview of the global distribution of autism, as opposed to narrow snapshot of a specific region, future surveys should aim for more equal distribution of ethnicities with each ethnic group having similar sizes.

Then there are issues regarding the original dataset's source and format. The dataset originally contained many demographics features with several containing categorical string values. However, not all of the string values are consistently formatted as the ethnicity feature contains many different spellings of the same ethnicity name. A possible remedy is to have participants answer ethnicities using a screening method type integer with a number representing a specific ethnicity. This would ensure more consistency and would allow for easier translation of data values into machine learning. Furthermore, the Nationality feature contains too many countries and therefore causing too much interference; for this reason, the original Nationality feature was dropped in the Capstone study. An alternative would be to include a demographic feature of region (i.e. North America, Africa, etc.), which would add a geographic aspect that would contain fewer unique variables. These considerations would lead to better comparisons of demographics and avoiding biases.

Recommendations

1. More resources and ASD treatment should go to Black, Hispanic/Latino, and White European ethnicities. These ethnicities have the highest rates of ASD diagnosis and higher scores associated with increased ASD. Furthermore, while Black and Hispanic/Latino have smaller representation than White European in the original data set, their relatively high rates show that ASD play a significant role in these ethnicities.
2. More studies are needed on Hispanic/Latino and Black ethnicities. While these ethnicities have smaller representation than White European, they have similar ASD diagnosis rates and scores. As such, there could be further studies directly comparing these ethnicities to White/European, preferably with equal ethnic sample size.
3. Future research on Autism should prioritize ethnicity over gender. The Random Forest models showed that certain ethnicities seem to play a larger role in determining diagnosis rates and scoring than gender. Furthermore, ethnicities have significant differences in scoring unlike genders, and the disparity in diagnosis rates are more pronounced between ethnicities than gender.

Sources and Links:

https://www.kaggle.com/faizunnabi/autism-screening