Autistic Spectrum Disorder (ASD) Screening for Adults

Team 5
Abdeali Arsiwala, Anjali Arya, Queenie Chao, Bowen Su, Michelle Wen

Foster MSBA, University of Washington
BUS AN 516 A Wi 24: Operations Research Data Analytics

Professor Masha Shunko
March 3, 2024

Executive Summary:

The report focuses on early diagnosis of autism spectrum disorder (ASD) in adults to reduce healthcare costs. Its main goal is to develop strategies for early ASD diagnosis, emphasizing the importance of timely interventions. Utilizing a global dataset from ASD screening applications, the research encompasses data collection, preprocessing, and analysis using machine learning models. The key findings highlight the effectiveness of early diagnosis in managing ASD and its impact on healthcare resource allocation. This project is significant due to the high healthcare costs associated with ASD, particularly in undiagnosed adults. Early diagnosis enables more effective and cost-efficient interventions, reducing long-term care expenses. By enhancing existing screening methods and raising public awareness, the research contributes to a more efficient healthcare system, addressing the challenges of prolonged diagnosis periods and promoting informed decision-making in ASD management.

Table of Contents:

1. Introduction:

a. Project Description:

The project aims to address the significant healthcare costs associated with autism spectrum disorder (ASD) by focusing on early diagnosis as a means of cost reduction. Early identification of ASD conditions is crucial for achieving cost-effective healthcare, as it enables informed decisions on whether to pursue comprehensive assessments and interventions.

b. Research Purpose:

The primary objective of the research is to explore and develop strategies for early diagnosis of ASD, emphasizing its importance in reducing healthcare costs. By identifying potential markers or indicators, the research seeks to contribute to the enhancement of assessment models, ultimately leading to more effective and timely interventions for individuals with ASD.

c. Significance:

ASD is known to be linked to substantial healthcare costs, including expenses related to assessments, interventions, and long-term care. The significance of this research lies in its potential to contribute to cost reduction by promoting early diagnosis. Timely identification of ASD allows for informed decision-making regarding the necessity and scope of assessments and interventions, ultimately leading to more efficient resource allocation.

d. Motivation:

The project is motivated by the challenges posed by lengthy waiting times for ASD diagnosis. The extended waiting periods can hinder timely interventions, potentially impacting the overall effectiveness of treatments. By addressing these challenges, the research aims to contribute to reducing waiting times for ASD diagnosis and enhancing the overall efficiency of the healthcare system in managing ASD cases.

e. Application and Awareness:

The research findings will be utilized to enhance existing screening applications, such as ASD tests, by providing valuable insights and markers for early diagnosis exploring more effective ways to screen potential autistic patients based on data available. The goal is to improve the effectiveness of these applications, enabling them to play a crucial role in the early identification of ASD.

Additionally, the project aims to contribute to raising awareness about ASD through strategic marketing messages. By disseminating information about the importance of early diagnosis and the role of screening applications, the research seeks to increase awareness among the public, healthcare professionals, and relevant stakeholders.

2. Dataset Overview:

a. Source:

Fadi Fayez Thabtah

Department of Digital Technology

Manukau Institute of Technology, Auckland, New Zealand

Dataset Link: ASD Screening Data for Children

b. Data Collection:

The dataset used in this project is obtained through autism spectrum disorder (ASD) screening applications. Fadi Fayez Thabtah, affiliated with the Department of Digital Technology at the Manukau Institute of Technology in Auckland, New Zealand, appears to be the primary contributor or researcher responsible for collecting and curating this dataset.

c. Scope:

1. **Geographic Scope:**
   The dataset covers a global scale, incorporating patient data from diverse regions, countries, and continents. This wide geographical scope suggests that the dataset captures a variety of demographics, making it valuable for understanding ASD trends on a global scale.
2. **Thematic Scope:**

   The dataset falls within the thematic scope of medical, health, and social science. This implies that the data encompasses information relevant to medical and health-related aspects of ASD, as well as social science perspectives. The inclusion of data from multiple thematic areas may offer a comprehensive view of various factors associated with ASD.

   This dataset, collected through ASD screening applications, serves as the foundation for the research project described earlier. Its global and thematic scope allows for a broad analysis of ASD trends, contributing to the project's objectives related to early diagnosis, healthcare cost reduction, and the enhancement of assessment models. Researchers can leverage this dataset to identify patterns, markers, and factors influencing ASD, ultimately informing strategies for more effective screening and intervention.

3. Data Exploration and Methodology:

a. Variable Descriptions:

Variable	Data Type	Example	Description
Age	Number	25	Represents the age of the individual in years.
Gender	String	Male, Female	Indicates the gender of the individual, categorized as Male or Female.
Ethnicity	String	Latino, Asian	Describes the ethnicity of the individual, specifying cultural or regional background.
Born with Jaundice	Boolean	Yes, No	Indicates whether the individual was born with jaundice, a condition characterized by yellowing of the skin and eyes.
Family Member with PDD (Pervasive Developmental Disorder)	Boolean	Yes, No	Specifies whether the individual has a family member with PDD, a group of developmental disorders that affect communication and behavior.
Q1 – Q10	Binary	0, 1	Binary responses to a set of ten questions (Q1 to Q10), where 0 may represent one response (e.g., No) and 1 may represent another response (e.g., Yes).
Relation	String	Self, Relative, Health Care Professional	Specifies the relationship of the individual to the survey or assessment, indicating whether the response is from the individual, a relative, or a healthcare professional.
Autistic	Boolean	Yes, No	Indicates whether the individual is diagnosed as autistic or not based on the dataset.

b. Descriptive Statistics:

1. Age:

The Mean Age of individuals is 29.7 years.

The Mean Age of individuals is 27 years.

2. **Gender Distribution:**

   There are 52% Male individuals and 48% Female individuals.

3. **Jaundice:**

   10% of the individuals in the dataset were born with jaundice.

4. **Family History (PDD - Pervasive Developmental Disorder):**

   13% of individuals in the dataset have a relative with PDD.

5. **ASD Traits (Initial Screen):**

   27% of individuals in the dataset exhibit traits associated with autism spectrum disorder (ASD) during the initial screening.

   These descriptive statistics provide a snapshot of key characteristics within the dataset. The information on age, gender distribution, jaundice occurrence at birth, family history of PDD, and the prevalence of ASD traits after the initial screening contributes to a better understanding of the demographic and health-related aspects of the individuals in the dataset. Researchers and healthcare professionals can use these statistics as a basis for further analysis and to draw insights into the factors associated with ASD.

4. Data Cleaning and Preprocessing:

a. Drop Uninformative Column:

The column *'age_desc'* was dropped as it contained the same value for all rows, providing no useful information.

b. Handling Categorical Variables:

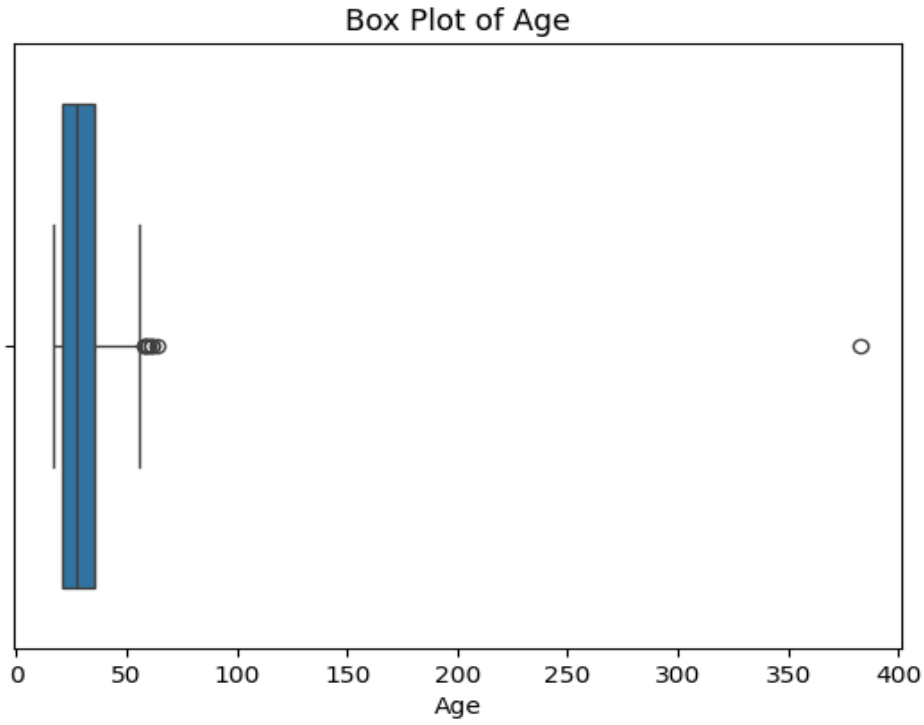
Categorical variables such as *gender, jundice, relative_autism, used_app_before*, and *Autistic* were encoded into numerical format for ease of analysis and model compatibility. Ethnicity and relation columns were standardized by grouping similar categories.

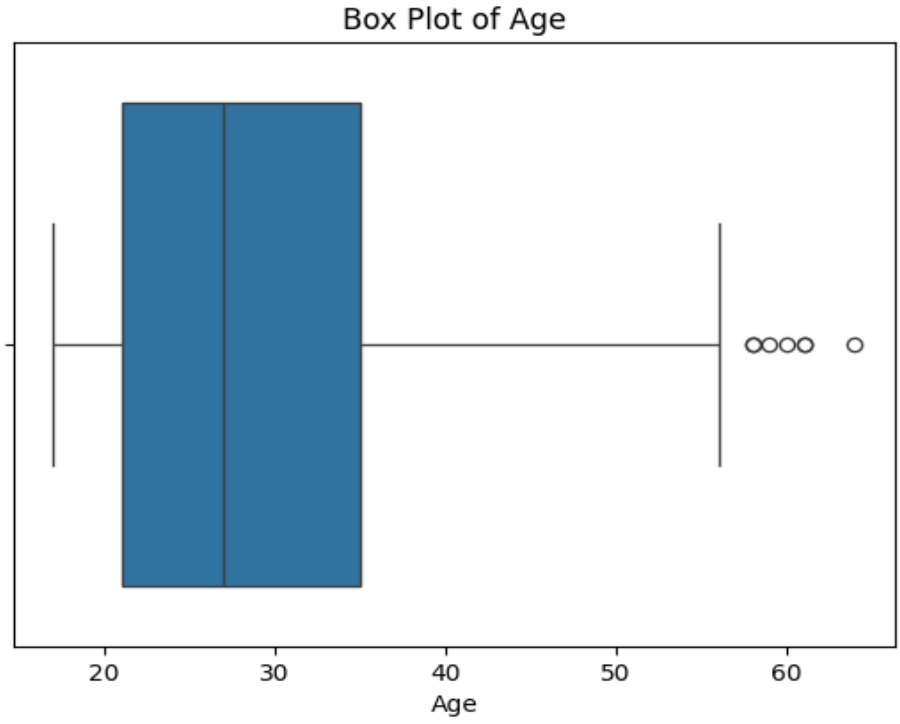
c. Handling Null Values:

1. Null values in the *'age'* column were identified as '?' values were replaced with NaN. Null values were then filled with the median age.
2. Null values in *'relation'* and *'ethnicity'* were put into 'Others' category.

d. Handling Outliers:

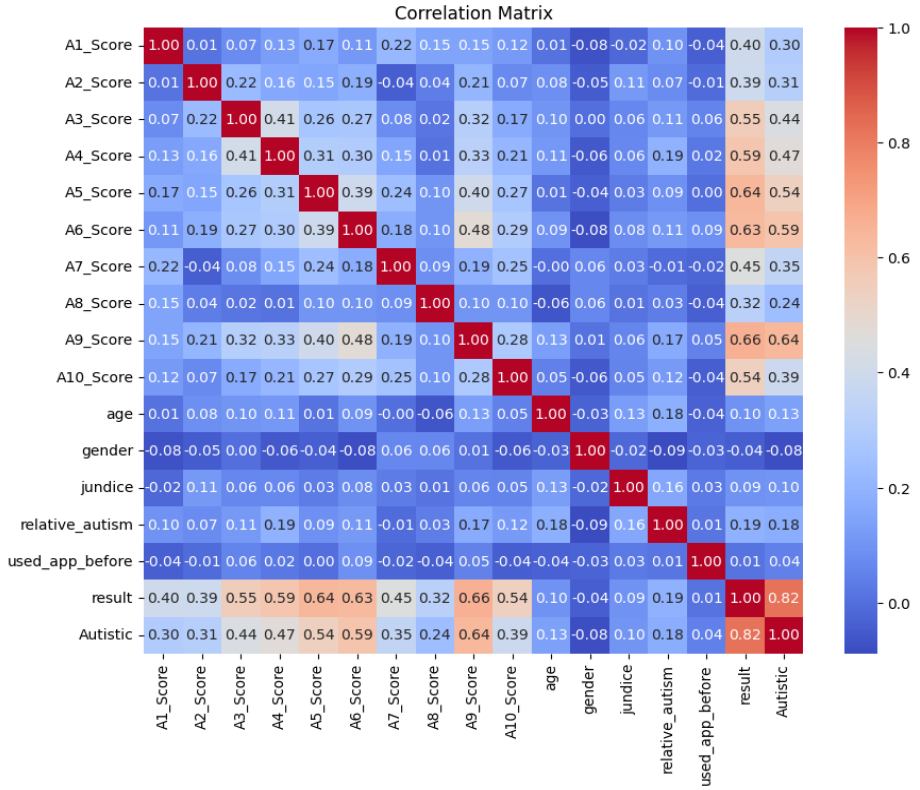
An outlier in the 'age' column (with age 383) was identified and replaced with the median age to mitigate its impact on the analysis.





e. Remove Redundant Column:

The column *'result'* was dropped to avoid multicollinearity, as it showed high correlation with variables A1 to A10 (sum of A1 to A10).



f. Categorizing Countries:

Country names with counts less than 20 were identified and replaced with 'Other' to reduce granularity and enhance analysis.

```
United States            113
United Arab Emirates      82
New Zealand               81
India                     81
United Kingdom            77
                         ...
China                      1
Chile                      1
Lebanon                    1
Burundi                    1
Cyprus                     1
Name: contry_of_res, Length: 67,
```

```
Other                    196
United States            113
United Arab Emirates      82
New Zealand               81
India                     81
United Kingdom            77
Jordan                    47
Australia                 27
Name: contry_of_res, dtype: int64
```

g. Relation Standardization:

The *'relation'* column was standardized by combining similar categories, such as grouping *'Health care professional'* with 'Others'.

h. One-Hot Encoding:

One-hot encoding was performed on categorical variables *'ethnicity'*, *'contry_of_res'*, and *'relation'* to create binary columns for each category.

```
DataFrame Columns:
A1_Score, A2_Score, A3_Score, A4_Score
A5_Score, A6_Score, A7_Score, A8_Score
A9_Score, A10_Score, age, jundice
relative_autism, Autistic, ethnicity_Black, ethnicity_Latino
ethnicity_Middle Eastern, ethnicity_Others, ethnicity_South Asian, ethnicity_White-European
contry_of_res_India, contry_of_res_Jordan, contry_of_res_Other, contry_of_res_United Arab Emirates
contry_of_res_United Kingdom, contry_of_res_United States, relation_Parent, relation_Relative
relation_Self
```

i. Mean Centering Age:

1. Initially, VIF (Variance Inflation Factor) was calculated for each numeric feature in the Data Frame, excluding the target variable 'Autistic'. This step provides an indication of multicollinearity among the features.
2. The mean of the 'age' column was calculated, and the column was mean centered by subtracting the mean from each value. This centering process aims to mitigate multicollinearity and improve the stability of regression models.
3. After mean centering the 'age' column, VIF was reduced from 10.79 to 1.22.

	feature	VIF
0	A1_Score	4.159481
1	A2_Score	2.192981
2	A3_Score	2.511525
3	A4_Score	2.807811
4	A5_Score	2.846781
5	A6_Score	2.164798
6	A7_Score	2.065918
7	A8_Score	3.101283
8	A9_Score	2.337304
9	A10_Score	2.921689
10	age	10.790420
11	gender	2.220200
12	jundice	1.206059
13	relative_autism	1.354107
14	used_app_before	1.069078
15	ethnicity_Black	1.637102
16	ethnicity_Hispanic	1.306014
17	ethnicity_Latino	1.376712
18	ethnicity_Middle Eastern	2.970744
19	ethnicity_Others	3.519953
20	ethnicity_Pasifika	1.168327
21	ethnicity_South Asian	1.321904
22	ethnicity_Turkish	1.100881
23	ethnicity_White-European	4.527910
24	contry_of_res_India	3.310886
25	contry_of_res_Jordan	2.419313
26	contry_of_res_New Zealand	2.999548
27	contry_of_res_Other	5.353253
28	contry_of_res_United Arab Emirates	3.529530
29	contry_of_res_United Kingdom	3.063995
30	contry_of_res_United States	4.045012
31	relation_Parent	2.600037
32	relation_Relative	1.921336
33	relation_Self	14.974479

	feature	VIF
0	A1_Score	4.167501
1	A2_Score	2.192373
2	A3_Score	2.511588
3	A4_Score	2.806745
4	A5_Score	2.851625
5	A6_Score	2.164655
6	A7_Score	2.067807
7	A8_Score	3.110716
8	A9_Score	2.339927
9	A10_Score	2.923771
10	age	1.223203
11	gender	2.219063
12	jundice	1.204612
13	relative_autism	1.348948
14	used_app_before	1.070062
15	ethnicity_Black	1.621176
16	ethnicity_Hispanic	1.296907
17	ethnicity_Latino	1.375633
18	ethnicity_Middle Eastern	2.964455
19	ethnicity_Others	3.271082
20	ethnicity_Pasifika	1.171598
21	ethnicity_South Asian	1.323087
22	ethnicity_Turkish	1.095375
23	ethnicity_White-European	4.310474
24	contry_of_res_India	3.106385
25	contry_of_res_Jordan	2.295797
26	contry_of_res_New Zealand	2.736040
27	contry_of_res_Other	4.915425
28	contry_of_res_United Arab Emirates	3.472356
29	contry_of_res_United Kingdom	2.842530
30	contry_of_res_United States	3.802270
31	relation_Parent	2.440585
32	relation_Relative	1.860769
33	relation_Self	13.802348

5. Analysis and Findings:

a. Research Question:

Our study aimed to see if combining information about how someone behaves, their medical history, and their personal details could help us better guess if they might have signs of ASD. We wanted to find out if looking at all these different kinds of information together could tell us if someone needs more tests to see if they have ASD. We also wanted to figure out which specific things about a person's behavior, medical history, or personal details were most important for predicting ASD signs.

b. Methodology:

1. **Screen Significant Variables**:
   Initially, we employed the Chi-square test as a preliminary step to identify which variables were statistically significant and should be included in our predictive models.

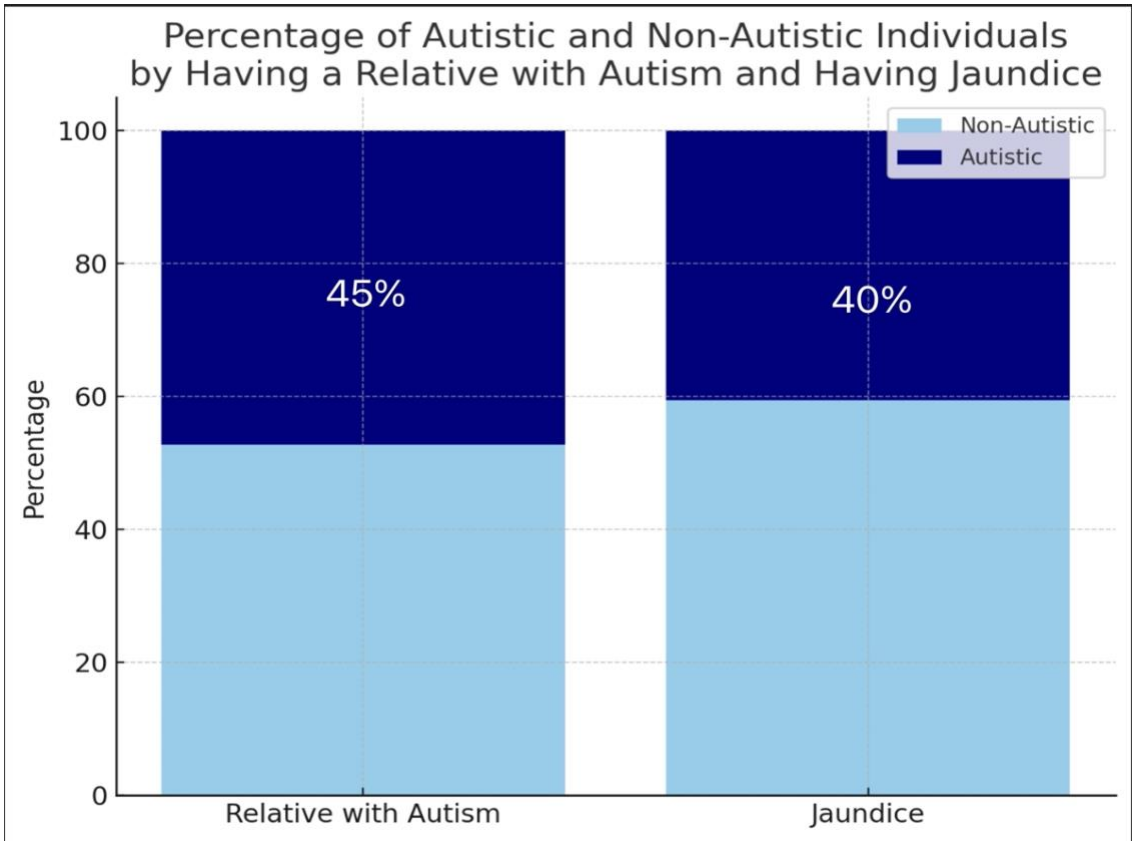
2. **Classification Model Building**:
   Following the variable selection process, we proceeded to construct two distinct types of models: CART (Classification and Regression Trees) and Logistic Regression. By building both types of models, we aimed to compare their performance and assess their suitability for predicting ASD traits.

3. **Choose Optimal Threshold for Each Model**:
   To optimize the performance of our models, we utilized Precision-Recall curves to select the optimal threshold for classification. This involved assessing the trade-off between precision and recall at various decision thresholds.

4. **Compare and Choose the Best Model**:
   Finally, we compared the performance of the CART and Logistic Regression models based on a set of predetermined evaluation metrics, such as accuracy and sensitivity. By rigorously evaluating each model against these metrics, we were able to determine which approach yielded the best predictive performance and was most suitable for our research objectives.

c. Chi Square Logistic Regression:

Exploring the significant relationship between the variables 'Family has Autistic Member,' 'Jaundice,' and individuals who are autistic from the dataset, we aimed to use the limited data at our disposal to provide actionable recommendations. However, certain variables, such as specific race, country of origin, age, or gender, are not immediately actionable. Yet, we identified two variables—having a family member with autism and being born with jaundice—that we can explore to potentially provide recommended decisions.



Our initial data visualization revealed that 45% of individuals with a family member with autism and 40% of those born with jaundice are autistic, suggesting a worthwhile exploration of their significant association.

Given that all three variables are Boolean, we utilized Chi-Square and Logistic Regression analyses to examine their relationships.

1. **Chi-square analysis on relative autism and Autistic:**

    a. Chi-square Statistic: 20.981873499582953
    b. P-value: 4.63649391560665e-06

2. **Chi-square analysis on Jaundice and Autistic:**

    a. Chi-square Statistic: 6.591397768445308
    b. P-value: 0.010247268414204275

    Chi-Square Analysis: Both jaundice and having a relative with autism showed significant associations with autism. The chi-square statistics and p-values underscored these relationships, indicating a noteworthy link.

    Logistic regression analysis on relative_autism, Jaundice in relations with Autistic

```
Confusion Matrix:
[[153   2]
 [ 53   4]]

Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.99      0.85       155
           1       0.67      0.07      0.13        57

    accuracy                           0.74       212
   macro avg       0.70      0.53      0.49       212
weighted avg       0.72      0.74      0.65       212

Model Coefficients (for relative_autism and jundice):
[[0.83762253 0.49540889]]
```

3. **Logistic Regression Analysis:**
   The coefficients for having a relative with autism and jaundice were 0.8376 and 0.4954, respectively, highlighting how changes in these predictor variables are associated with the log odds of being autistic. However, the model demonstrated limited ability in predicting autistic cases based solely on these two variables, as reflected by the low recall for autistic cases.

4. **Recommendation:**
   These findings can be instrumental in crafting marketing messages and raising awareness, encouraging more individuals to take Professor Fadi Fayez Thabtah's ASD screening application. Suggested ad copy questions include "Do you have a family member that is Autistic?" or "Were you born with Jaundice?" to attract those more likely to be autistic to undergo assessment and seek necessary treatment.

   Furthermore, we recommend that healthcare policymakers and insurance companies consider these relationships in a broader dataset to inform their actions.

d. CART Model:

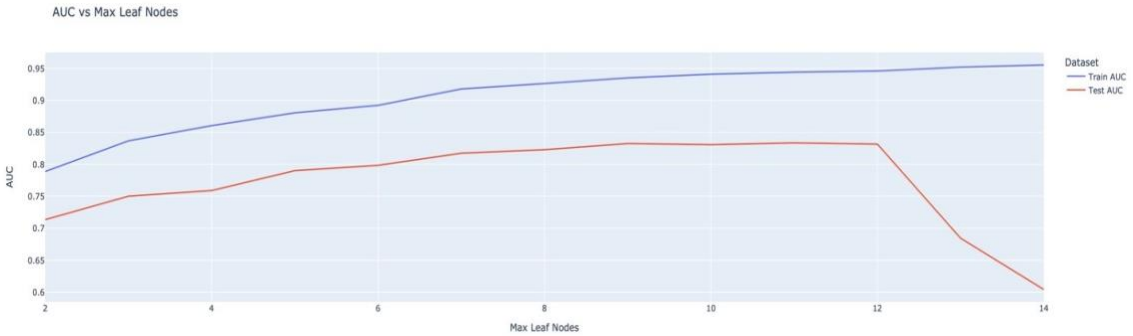
1. **Why CART?**
   CART (Classification and Regression Trees) models were selected for their ability to capture non-linear relationships between predictor variables and the target variable, which is crucial in the context of autism classification. In autism diagnosis, there may exist intricate, non-linear interactions among various features, and CART models excel at identifying and representing such complex patterns.

2. **CART Assumptions**:
   CART models do not assume a specific functional form for the relationships between predictor variables and the target variable, allowing for flexibility in modeling diverse data patterns. CART models operate on the assumption that relationships between predictor variables and the target variable can be effectively captured through hierarchical splits.

3. **Variables Overviews**:

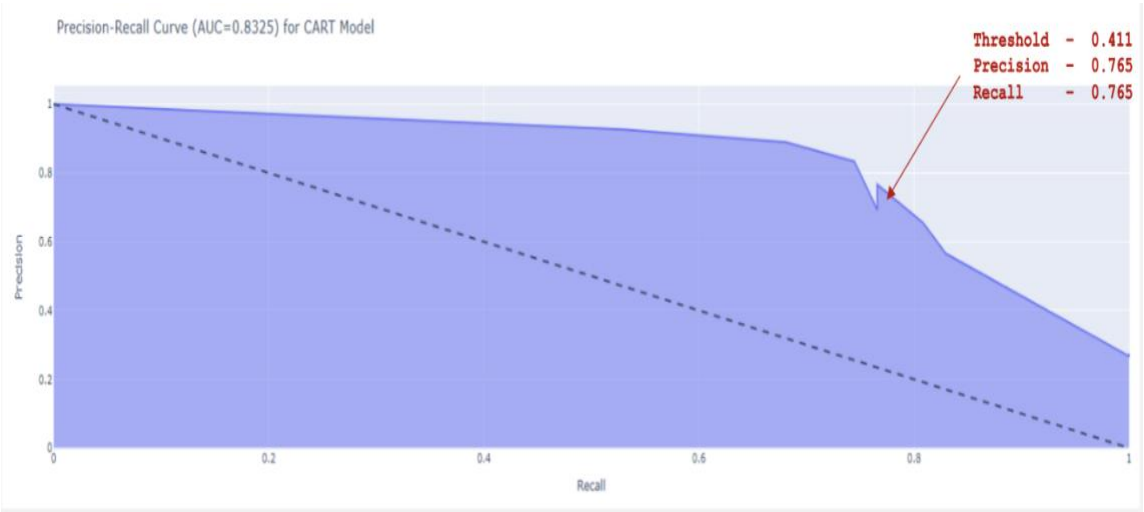
Data Types	Variables	Note
Numerical	Age	
Categorical	Question Score	A1_Score~A10_Score
	Jaundice	
	Ethnicity	Middle Eastern, South Asian, White-European, Others
	Country of Residence	India, Jordan, United Arab Emirates, United Kingdom, United States, Others
	Relation	Self, Parent, Relative, Others

4. **Maximum Leaf Nodes Decision**:
   To establish the maximum number of leaf nodes, we constructed an AUC graph. Opting for the AUC curve over the ROC curve was intentional as our emphasis was on sensitivity (recall) in this scenario. Our primary goal was to reduce the occurrence of false negatives, especially concerning late autism diagnoses. Upon analysis, we noted a substantial decline in the AUC of the test dataset when the maximum leaf nodes surpassed 12. Hence, to prevent overfitting, we capped the maximum leaf nodes at 12.

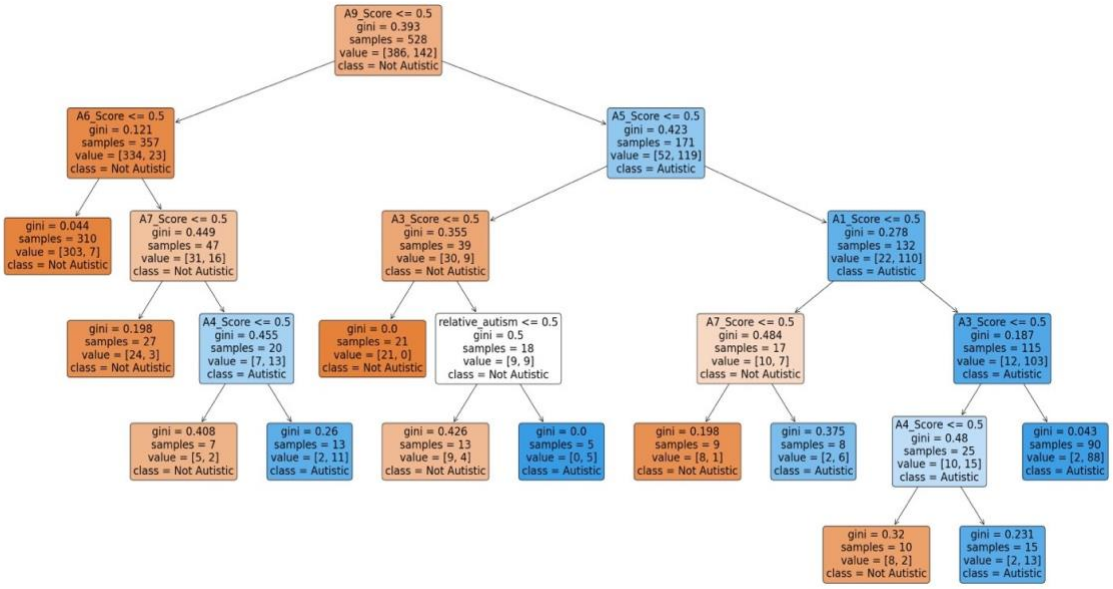


5. **Threshold Decision**:
   We utilized the AUC curve to identify our optimal threshold, determined to be 0.411. This threshold was selected based on its ability to optimize our model's performance by striking

a balance between sensitivity and specificity, thereby enhancing the accuracy of ASD trait classification.



6. **Model Overview**:



- We noticed that the CART model mainly splits based on the 10 questions about behavior related to autism. This means these questions are important for figuring out if someone might have autism.
- The Gini value, which shows how good our model is at telling if someone has autism, was 0.5. That's the worst score possible, which suggests our model isn't good at using things like demographic or medical history to predict autism.
- In conclusion, even though the CART model is good at understanding behavior, it's not so great at using other information like age or medical history to predict autism. This tells us

we might need to try different ways or include more information to make better predictions about autism.

e. Logistic Regression:

1. **Logistic Regression Assumptions**:
   The assumptions of logistic regression are pertinent for autism screening models as they facilitate understanding the binary outcome and the linear relationship of predictors, aiding in effective screening and intervention strategies. Additionally, logistic regression enables the determination of variable weights and importance, providing insight into the relative contributions of factors like age, ethnicity, and family history to autism occurrence.

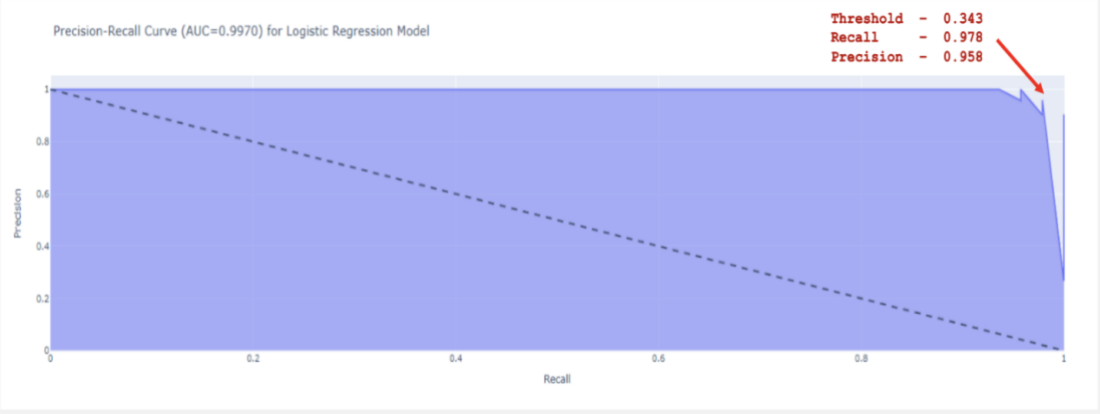
2. **Model Overview**:

```
l number of coefficients: 28
```

Variable	Coefficient
A9_Score	2.434209
A5_Score	2.107205
A6_Score	2.066951
A7_Score	2.021794
A4_Score	1.970274
A1_Score	1.938328
A10_Score	1.806594
A3_Score	1.779005
A8_Score	1.73359
A2_Score	1.492332
ethnicity_White-European	0.685707
ethnicity_Others	0.424538
contry_of_res_Other	0.401889
ethnicity_Black	0.350574
relative_autism	0.325547
contry_of_res_United States	0.308736
jundice	0.251837
relation_Self	0.192002
:ontry_of_res_United Arab Emirates	0.190385
ethnicity_Latino	0.100852
age	-0.002344
ethnicity_South Asian	-0.02111
relation_Parent	-0.030304
contry_of_res_India	-0.056687
relation_Relative	-0.324825
contry_of_res_Jordan	-0.350354
contry_of_res_United Kingdom	-0.393579
ethnicity_Middle Eastern	-0.415436

- We noticed that the variables linked to behavioral questions have larger coefficients compared to other variables in our model. This means that behaviors related to autism play a crucial role in predicting autism traits compared to other factors we looked at.
- Additionally, our model can consider other variables like certain ethnicities, whether someone had jaundice, and if they have relatives with ASD. This suggests that these factors also contribute to our model's ability to predict autism traits, although to a lesser extent compared to behavioral variables.

3. **Threshold Decision**:

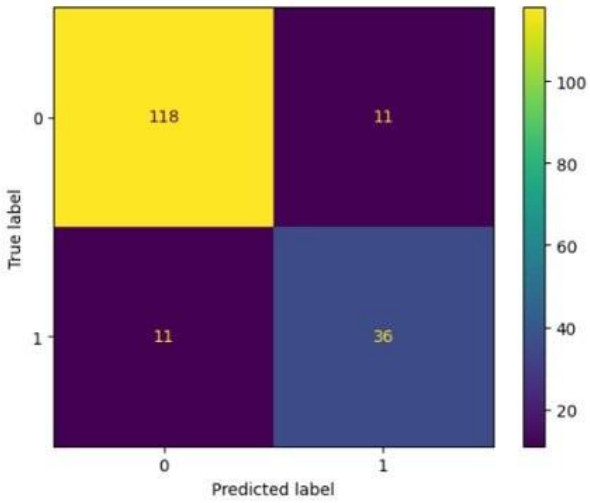
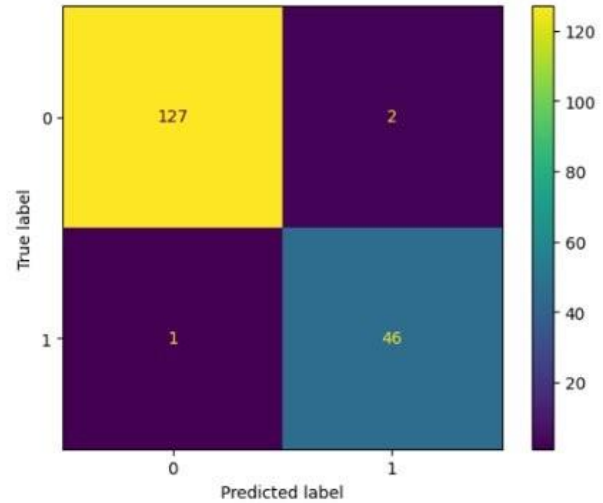


To optimize recall while balancing precision in logistic regression, we selected a final threshold of 0.343. This threshold was chosen to maximize the model's ability to correctly identify individuals with ASD traits, minimizing false negatives.

f. Confusion Matrix:

Based on sensitivity and accuracy metrics, logistic regression outperforms CART. Logistic regression achieves a sensitivity of 0.979, significantly higher than CART's sensitivity of 0.766. Similarly, logistic regression achieves an accuracy of 0.983, surpassing CART's accuracy of 0.875.

Logistic regression also offers the advantage of being able to incorporate demographic and medical information effectively into the predictive model. Unlike CART, which primarily relies on behavioral questions, logistic regression can capture the predictive power of demographic and medical variables. By including factors such as ethnicity, jaundice at birth, and family history of ASD, logistic regression can provide a more comprehensive understanding of the factors contributing to ASD traits. This capability enhances the model's ability to generalize predictions beyond behavioral patterns alone. Thus, logistic regression not only excels in predictive performance but also provides valuable insights into the influence of demographic and medical factors on ASD traits.

CART	Logistic Regression
	

6. Conclusion:

We analyzed data from Professor Fadi Fayez Thabtah's ASD screening applications, focusing on the relationships among 'Jaundice,' 'Family Member with Autism,' and 'Autism.' Our goal was to explore how early diagnosis could reduce costs by:

a. Raising Awareness with Detailed, Targeted Messages:

We suggest using ad copy that asks questions like "Do you have a family member who is Autistic?" or "Were you born with Jaundice?" This approach aims to attract individuals who may be more likely to be autistic, encouraging them to take the assessment and seek treatment if necessary.

b. Recommendations for Healthcare Policymakers and Insurance Companies:

We urge healthcare policymakers and insurance companies to examine whether there is a significant association between these three variables in a larger dataset. If such a relationship exists, it could serve as a basis for reaching out to those who match the description for assessments and treatment.

c. Enhancing Screening Applications with Additional Variables:

Professor Fadi Fayez Thabtah's ASD screening applications currently determine the need for further professional assessment based on responses to questions Q1 to Q10. Our logistic regression model introduces additional variables that may indicate the need for further assessment. With a new threshold of 0.343, lower than the current 0.6, and by incorporating demographic and medical condition data into the classification process, we believe our model could improve the screening process. This improvement aims to reduce false negative cases,

aiding those who need help but were not identified by the current threshold used in the ASD screening applications.

7. Source:

Thabtah,Fadi. (2017). Autistic Spectrum Disorder Screening Data for Children . UCI Machine Learning Repository. https://doi.org/10.24432/C5659W.