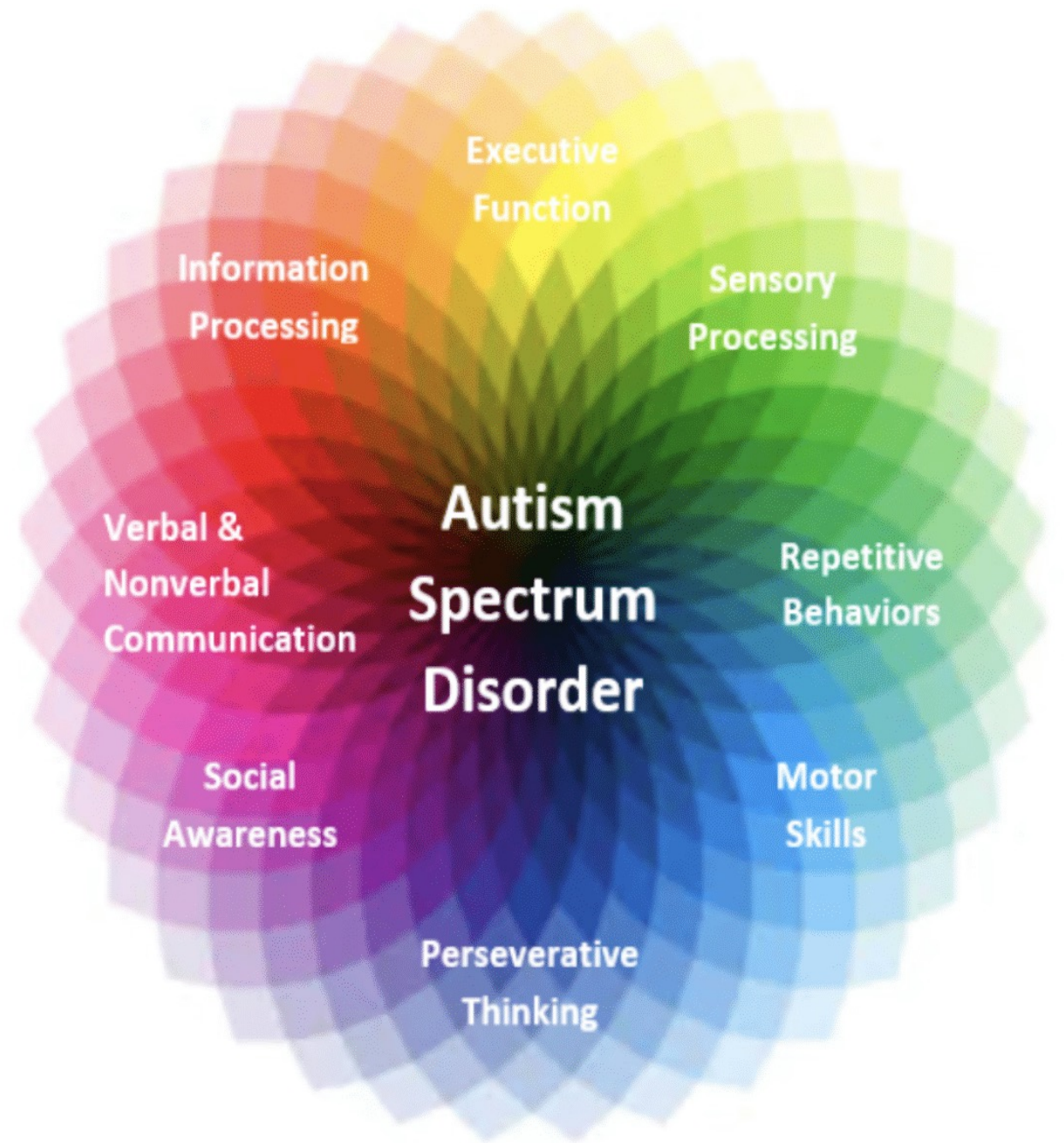


AUTISTIC SPECTRUM DISORDER (ASD) SCREENING FOR ADULTS

TEAM 5

- ABDEALI ARSIWALA
- ANJALI ARYA
- BOWEN SU
- QUEENIE CHAO
- MICHELLE WEN





Significance:

- ASD linked to substantial healthcare costs.
- Early diagnosis crucial for cost reduction: Timely identification of conditions is crucial for cost-effective healthcare, enabling informed decisions on whether to pursue a comprehensive assessment and intervention.



Motivation:

- Address challenges of lengthy waiting times for ASD diagnosis.
- Utilize our research finding to help Screening Application such as ASD test with marketing message, raising more awareness
- Enhance assessment model, increase effectiveness

DATASET OVERVIEW



Source:

Fadi Fayez Thabtah

Department of Digital Technology

Manukau Institute of Technology,

Auckland, New Zealand

<https://archive.ics.uci.edu/dataset/419/autistic+spectrum+disorder+screening+data+for+children>

- Data collected through ASD screening applications.

Scope:



Geographic Scope –

The dataset covers a global scale, incorporating patient data from diverse regions, countries, and continents.

Thematic Scope –

Medical, health and social science

DATASET DETAILS

Descriptive Statistics

- Mean age is 29.7, Median age is 27
- 52% Male, 48% Female
- 10% born with jaundice
- 13% relative had PDD
- 27% show traits of ASD (initial screen)

Variable	Data Type	Example
Age	Number	25
Gender	String	Male, Female
Ethnicity	String	Latino, Asian
Born with Jaundice	Boolean	Yes, No
Family Member with PDD	Boolean	Yes, No
Q1 – Q10	Binary	0, 1
Relation	String	Self, Relative, Health Care Professional
Autistic	Boolean	Yes, No

Q1-Q10

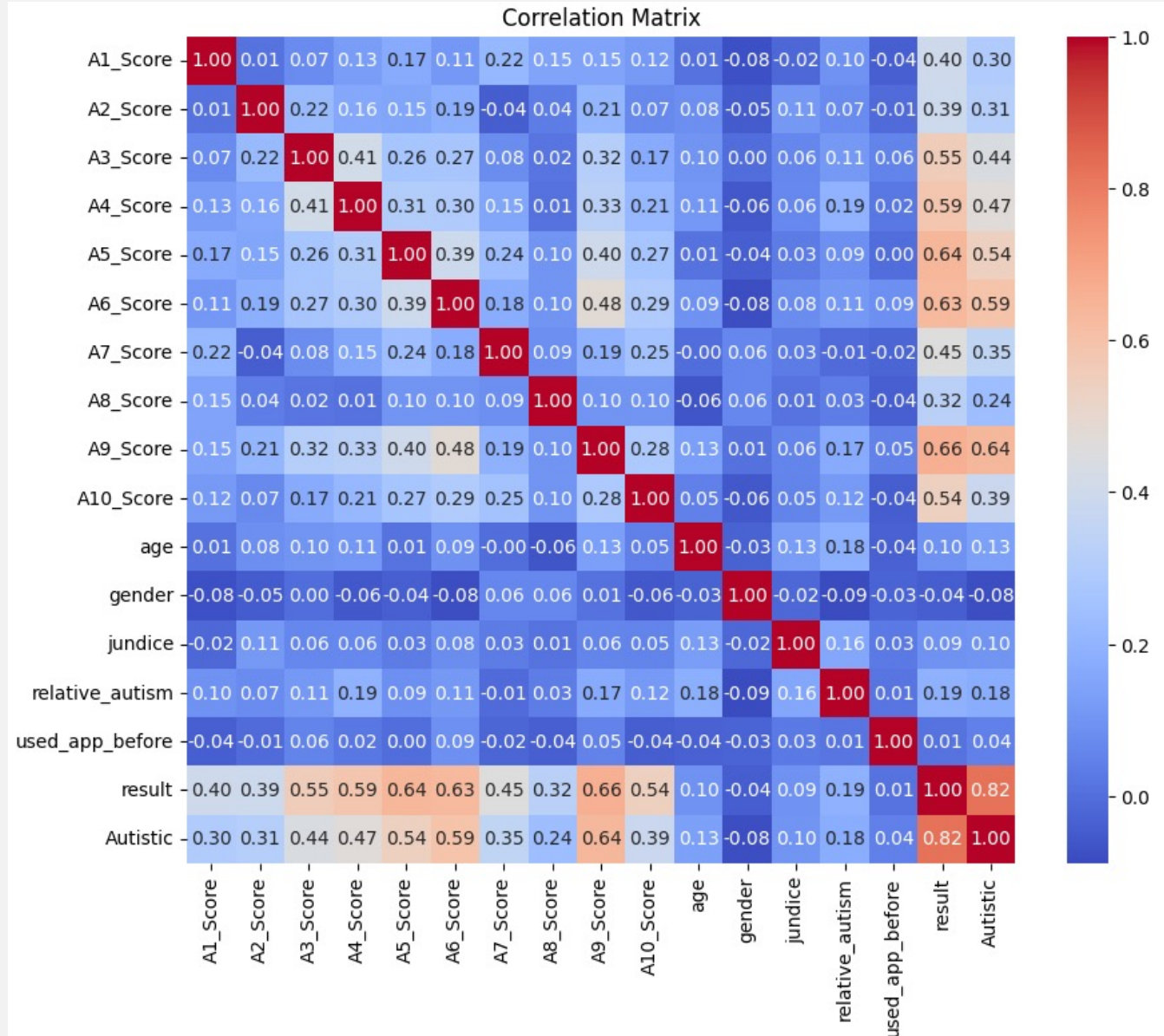
Please tick one option per question only:

		Definitely Agree	Slightly Agree	Slightly Disagree	Definitely Disagree
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				

DATA PREPARATION

Data Processing

- Removed the **redundant 'Result'** variable due to its high correlation with other variables.
- Encoded boolean features** (Gender, Jaundice, Relative Autism, Autistic) as binary (1/0).
- Grouped** countries of residence with a count less than 50 under the **'Others'** category.
- Grouped** 'Healthcare Professionals' in relation variable into the **'Others'** category.

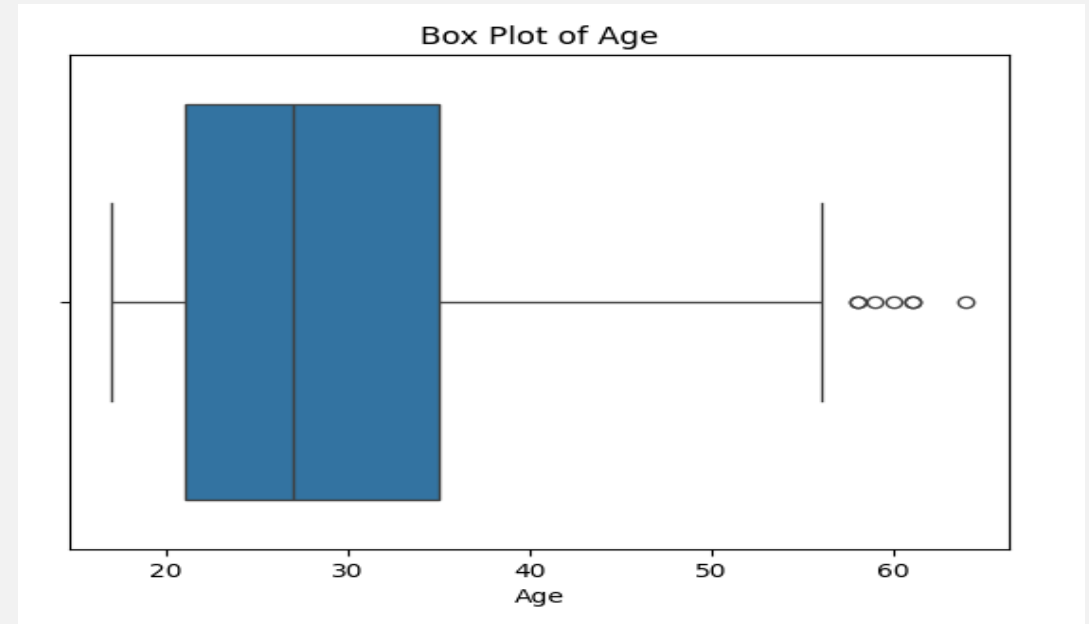
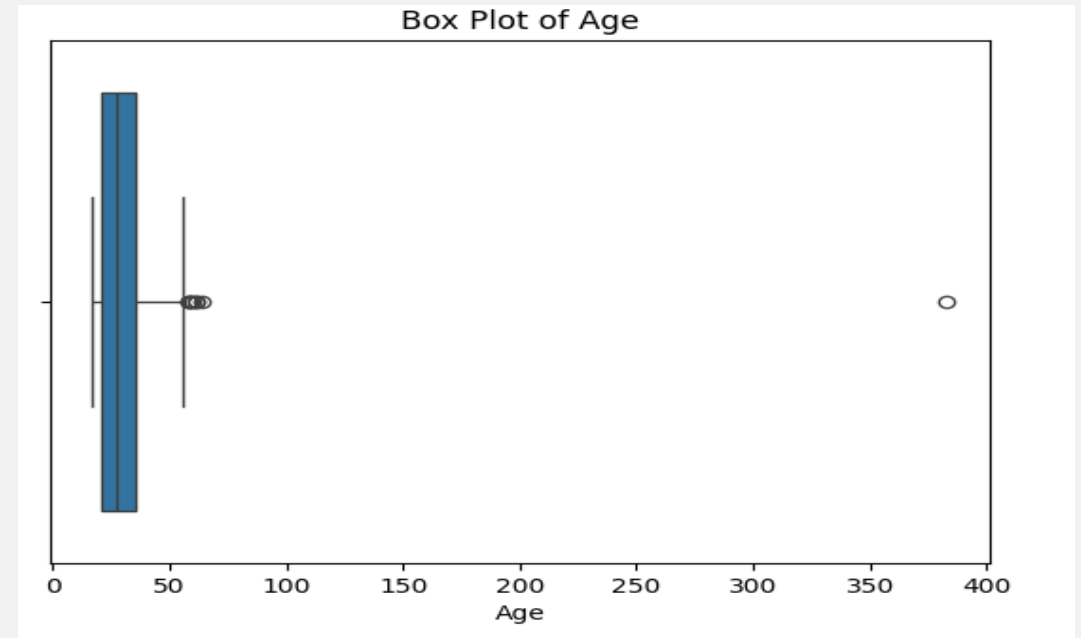


Missing Value Handling

- Missing 'Age' values imputed with median age.
- Missing 'Ethnicity' values added to 'Others' category.

Outlier Treatment

- Identified an outlier in 'Age' (value 383) likely due to reporting error (visualized via boxplot).
- Replaced outlier with median 'Age' value to reduce its influence on analysis.



VARIANCE INFLATION FACTOR

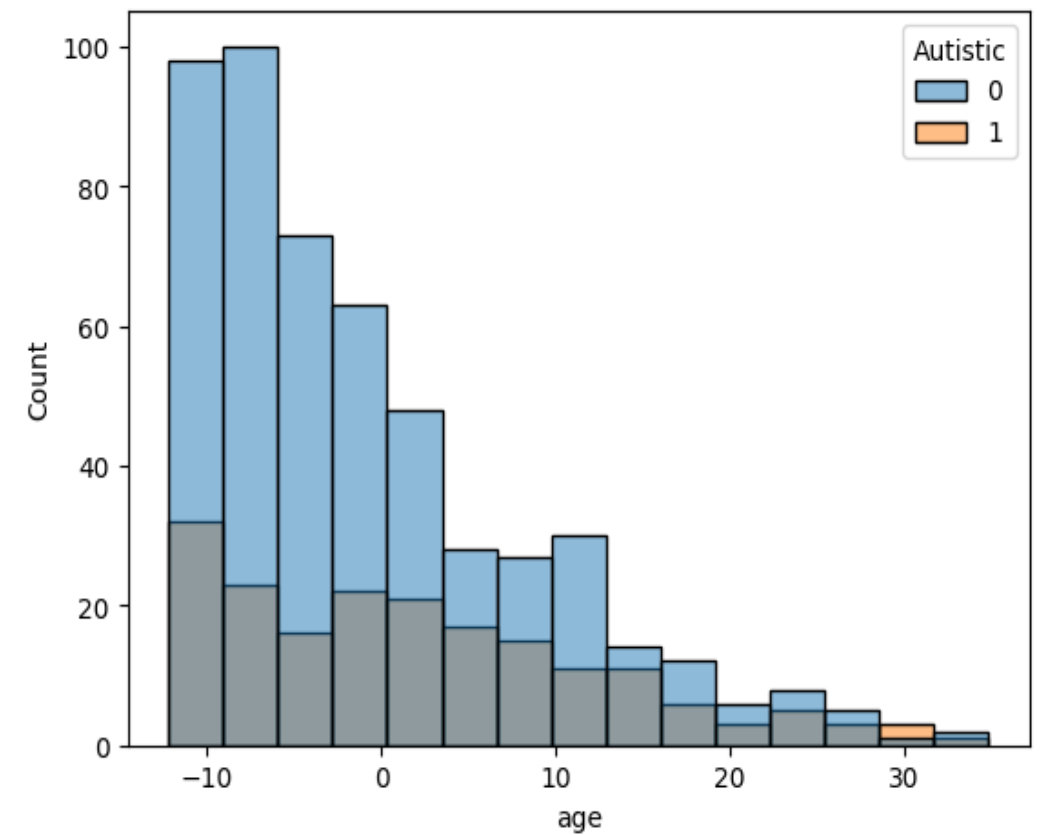
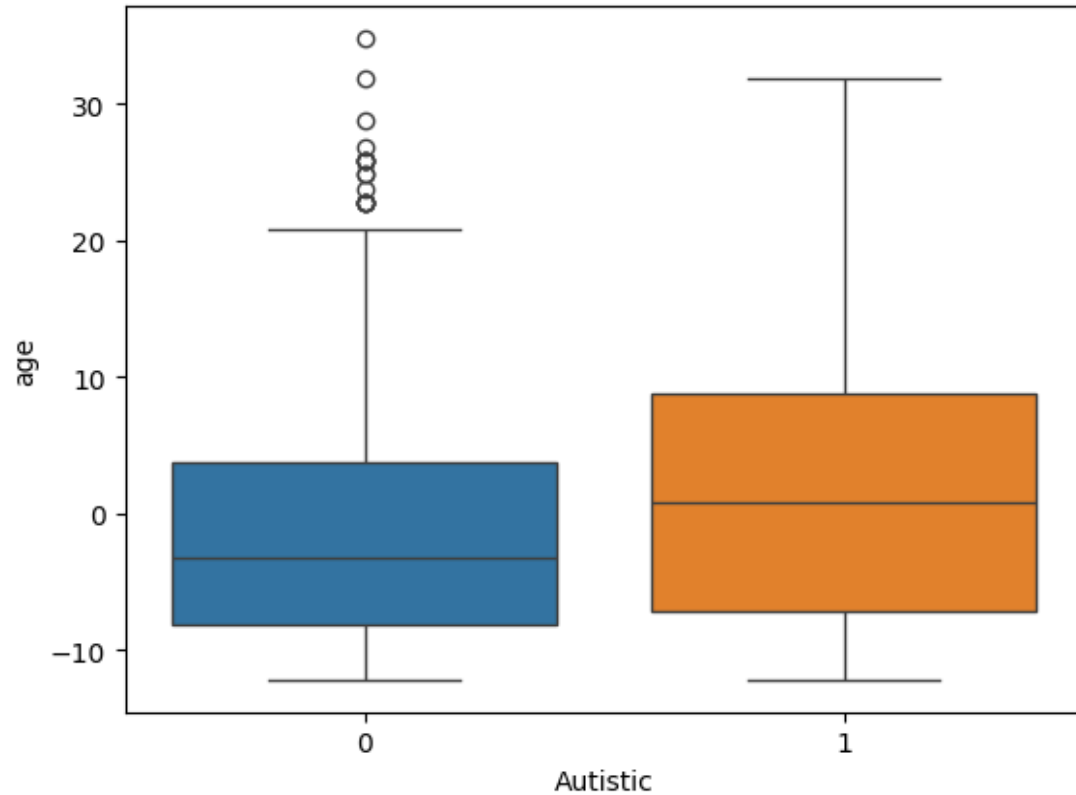
High VIF For age: 10.790

Therefore, we decided to perform mean centering.

VIF score is significantly reduced.

	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



1. The median age of people with autism is about the same as that of people without autism.
2. Also, the range of age for both values (Autistic and non- autistic) in the target class is almost the same.

Hence, age cannot be a reliable classifier for the target class.

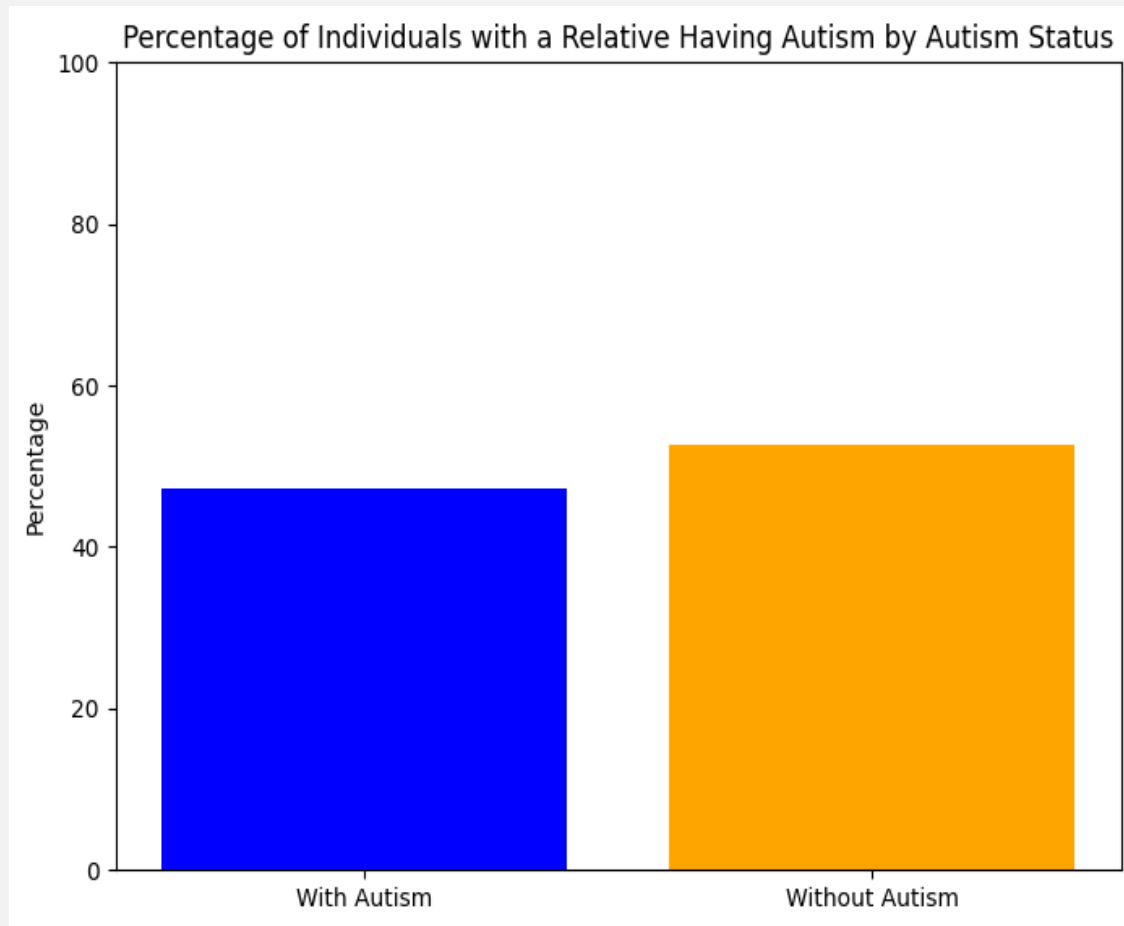
DEPENDENCE OF CATEGORICAL VARIABLES ON TARGET

DATASET DETAILS

Descriptive Statistics

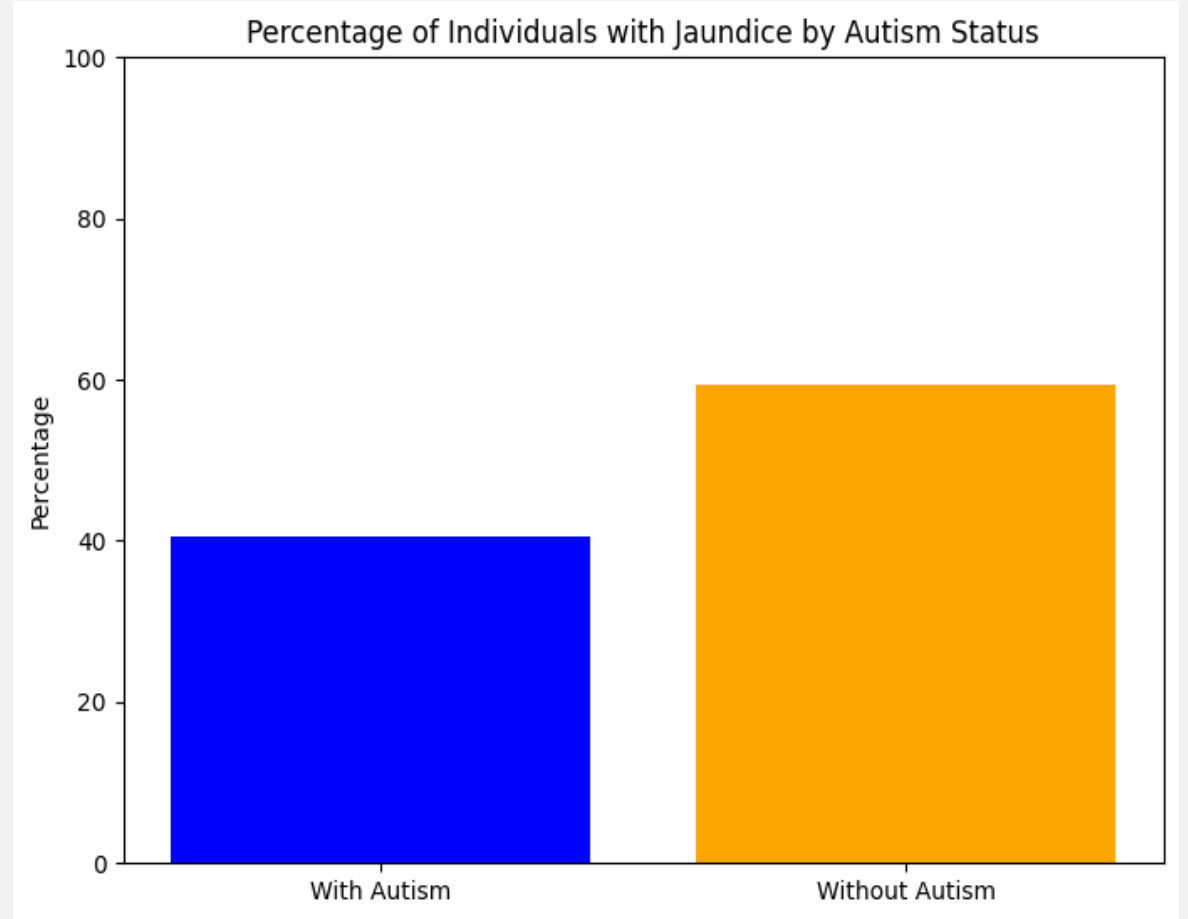
- Mean age is 29.7, Median age is 27
- 52% Male, 48% Female
- 10% born with jaundice
- 13% relative had PDD
- 27% needs additional assessment

Variable	Data Type	Example
Age	Number	25
Gender	String	Male, Female
Ethnicity	String	Latino, Asian
Born with Jaundice	Boolean	Yes, No
Family Member with PDD	Boolean	Yes, No
Q1 – Q10	Binary	0, 1
Relation	String	Self, Relative, Health Care Professional
Autistic	Boolean	Yes, No



Relationship between Family History & Autism:

- Roughly 45% of people having a "relative that is Autistic" were shown to have traits of ASD.



Relationship between Jaundice & Autism:

- 40% of people born with Jaundice were shown to have traits of ASD.

Relationship between ASD result and Family History.

- Chi square analysis results - 'relative_autism' and 'Autistic' :
 - Chi-square Statistic: 20.9819
 - P-value: < 0.0001

There is a **significant** association between the variables 'relative_autism' and 'Autistic' in the dataset, as indicated by the **p-value** being much less than **0.05**. This suggests that the presence of autism in relatives is associated with the need for full ASD assessment.

Relationship between ASD result Jaundice.

- Chi square analysis results - 'Jaundice' and 'Autistic' :
 - Chi-square Statistic: 6.59140
 - P-value: 0.0102

There is a significant association between the two variables, suggesting that individuals with jaundice are more likely to require additional autism assessment.

LOGISTIC REGRESSION ON ASD RESULT AND FAMILY HISTORY, AND JAUNDICE

- The coefficient for relative_Autism and Jaundice are 0.8376 and 0.4954.
- Specifically, having a relative with autism (relative_autism = 1) is associated with an increase in the log odds of needing additional assessment for autism, as indicated by the positive coefficient. Similarly, having jaundice (jundice = 1) is also associated with an increase in the log odds of needing additional assessment for autism, but its impact is smaller compared to relative autism based on the coefficient values.
- These results suggest that both having a relative with autism and having jaundice are associated with needing additional assessment for autism, with relative autism having a stronger association based on the model coefficients.

ADDITIONAL INFERENCES

- Similarly, **ethnicity_Pasifika**, **ethnicity_Hispanic**, **ethnicity_Turkish**, **country_of_res_New Zealand**, **relation_Parent**, and **relation_Relative** showed independence from the target variable in the chi-square test. Hence, we will not be generalizing our findings for these variables due to insufficient data.
- Consequently, these variables will be grouped into the 'Others' category for further analysis.

RECOMMENDATIONS

- The Study shows family has autistic patients, also people born with jaundice has significant association with the need for additional assessment for Autism.
- We recommend ASD testing App use this as marketing message to raise awareness, raising attention from people who are in this situation to take the ASD test.
- We recommend policy maker, insurance company to get more, related data and take action such as providing them access to assessment.

RESEARCH QUESTION

Question -

Use a person's behavioral data combined with medical and demographic data to predict whether the person shows traits of ASD.

Methodology -

- ✓ i. **Select statistically significant variables - {chi-square test}**
- ▶▶ ii. **Classification Model Building**
 - CART (Classification and Regression Trees) Model
 - Logistic Regression Model
- ▶▶ iii. **Choose Optimal Threshold for Each Model - (Precision-Recall Curve)**
- ▶▶ iv. **Compare and Choose the Best Model (based on metrics)**

MODEL BUILDING - CART

Why CART?

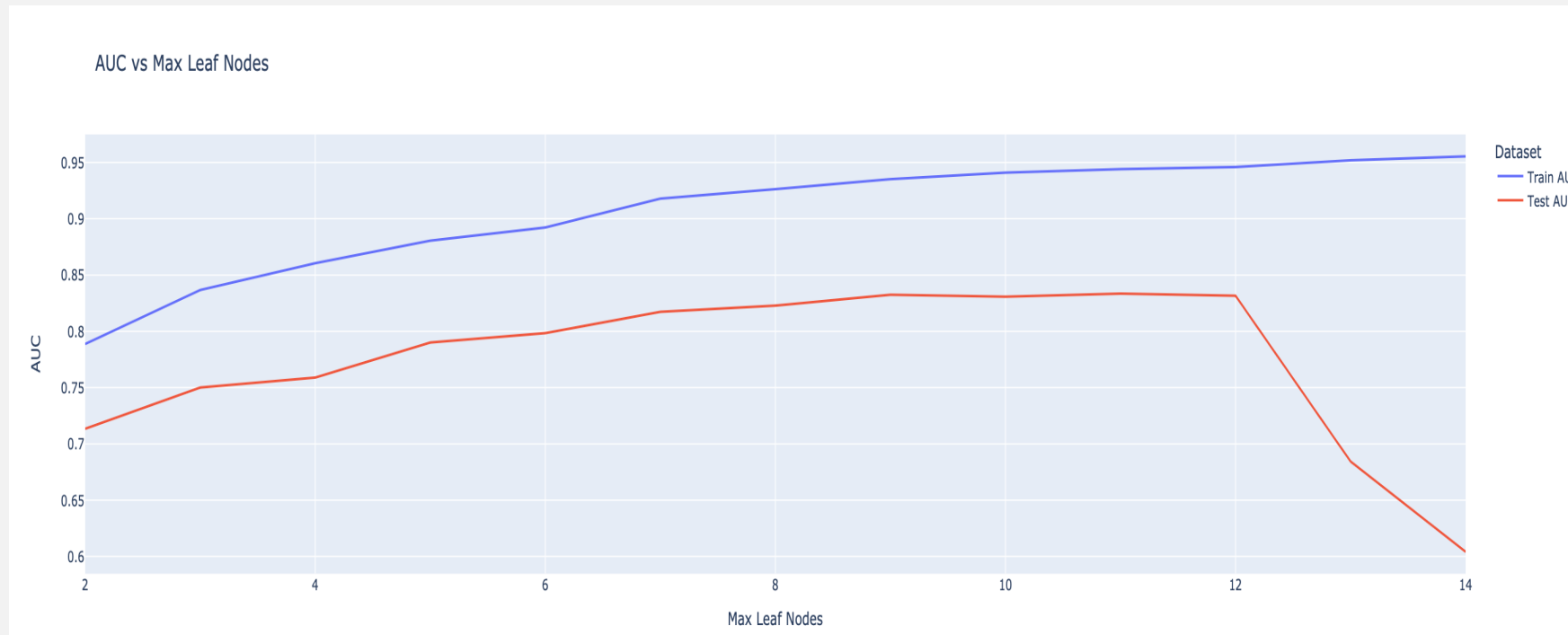
- Non-parametric and can capture non-linear relationships between predictor variables and the target variable.
- In the context of autism classification, there may be complex, non-linear interactions among various features, and CART models are well-suited to handle such patterns.

Assumptions:

Relationships between predictor variables and the target variable (Autistic) can be captured through hierarchical splits.

Data Types	Variables	Note
Numerical	Age	
Categorical	Question Score	AI_Score~AI0_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

PARAMETER TUNING

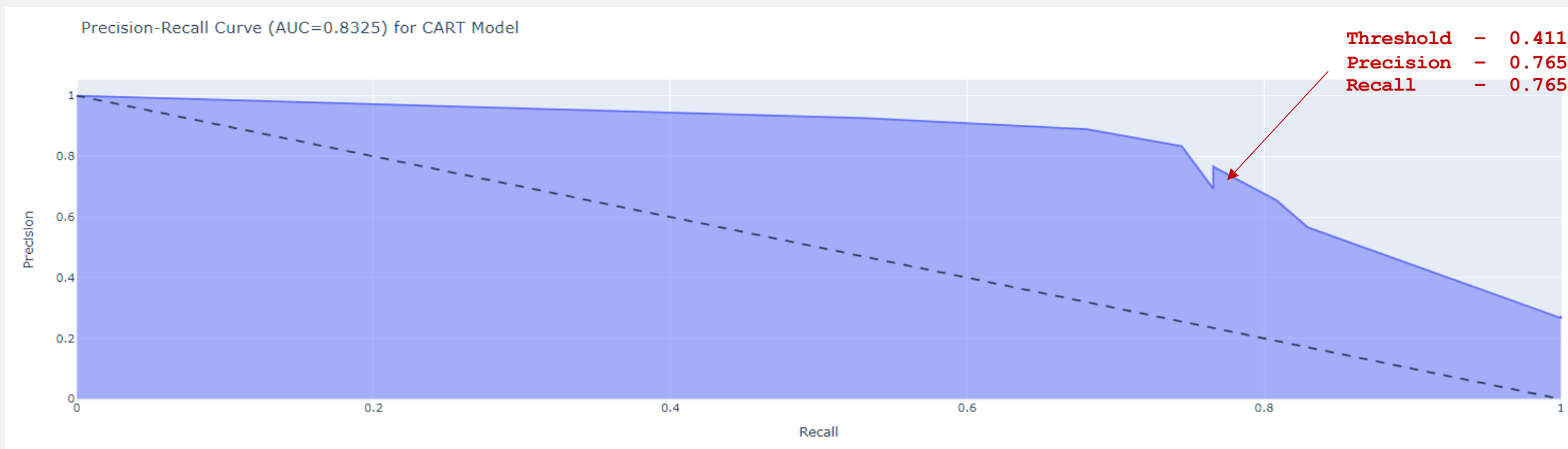


- Optimal AUC for both train and test at Max_Leaf_Nodes = 12
- No Overfitting

BEST THRESHOLD - (PRECISION-RECALL CURVE)

Why Precision-Recall Curve ?

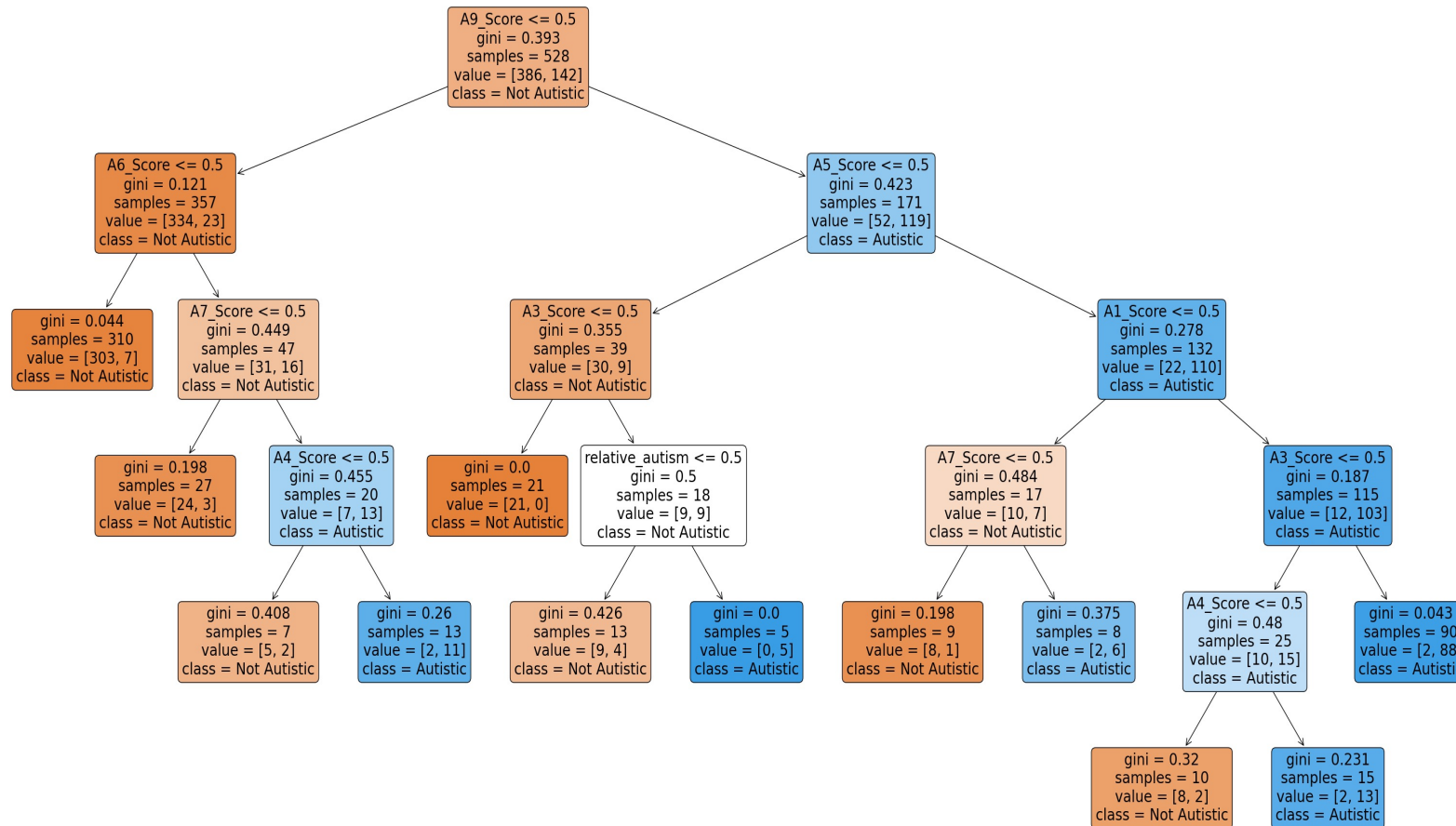
In case of imbalanced dataset where we have a large no of observations with negative values, i.e. 'Non-Autistic', it is beneficial to look at Precision-Recall curve, since Precision does not take into account the True Negatives, and hence is unaffected by the imbalance.



Goal -

- Maximize Recall (aka Sensitivity)
- While balancing Precision

MODEL VISUALIZATION - CART



- **Model Overview:**

- Decision nodes: A9, A5, A1 etc.
- Gini (relative_autism) = 0.5, which is the worst score.
- All decision nodes are derived from 10 behavioral questions. **CART is not able to generalize for demographic or medical information.**

- **Key Findings:**

- **(A9=0):** If someone finds it easy to understand what someone else is thinking by looking at their face, it's very likely that they are not autistic.

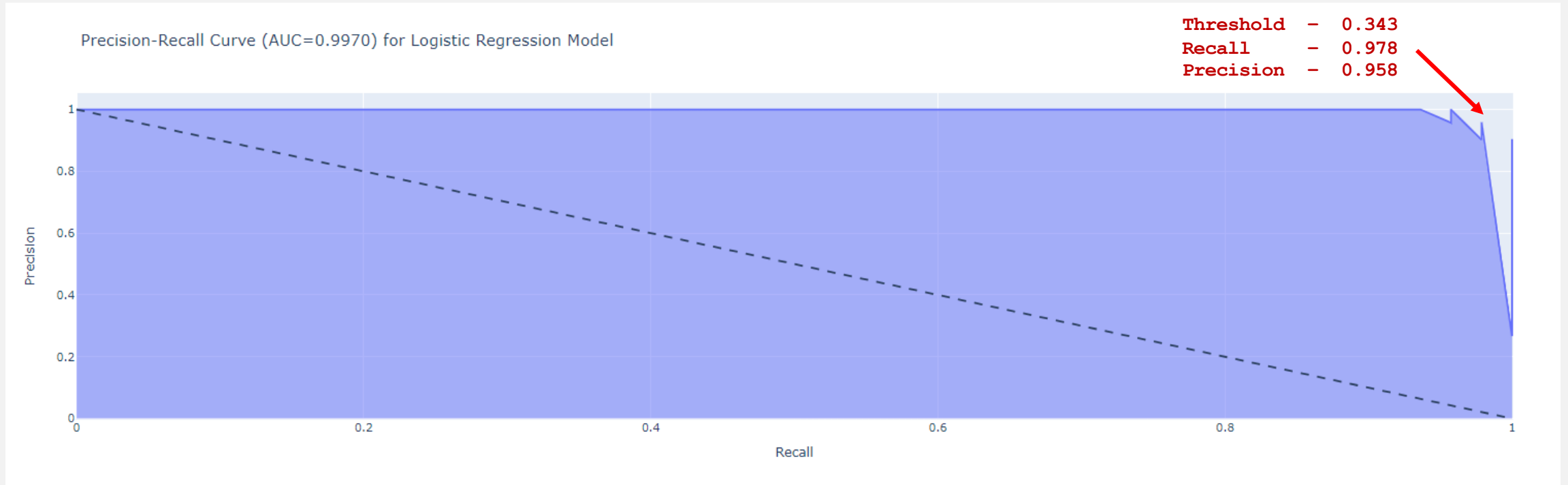
MODEL BUILDING - LOGISTIC REGRESSION

- Variables related to behavioral questions have a higher coefficient compared to the other variables.
- Our model is also able to account for other variables like **a few ethnicities, jaundice at birth, and relatives with ASD.**

1 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
jaundice	0.251837
relation_Self	0.192002
contry_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

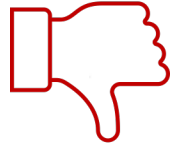
BEST THRESHOLD - (PRECISION-RECALL CURVE)



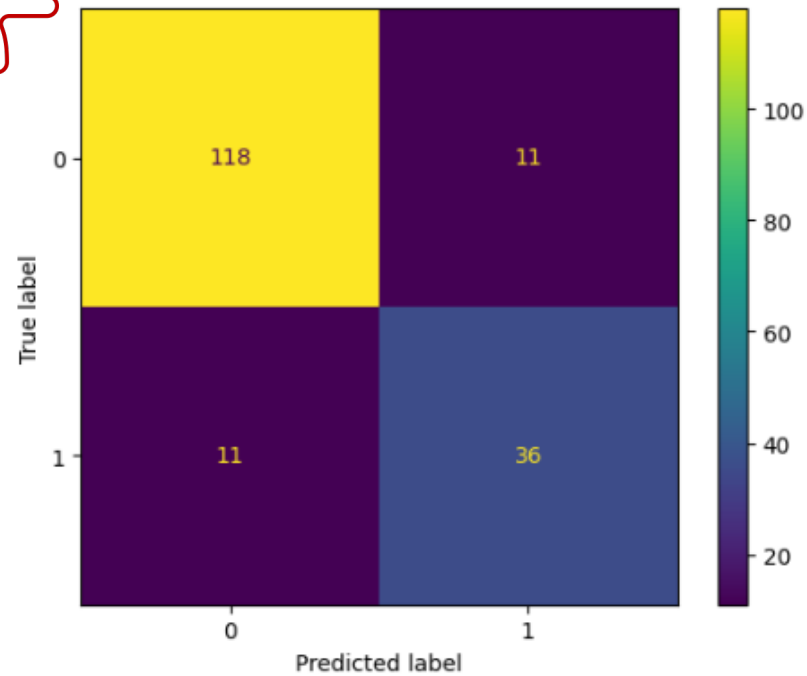
Goal -

- Maximize Recall (aka Sensitivity)
- While balancing Precision

FINDING BEST ALGORITHM



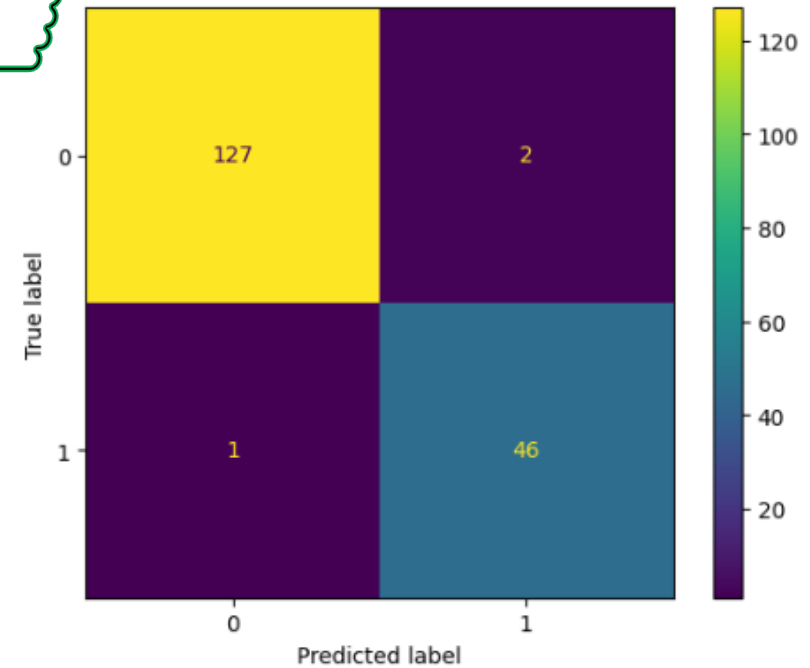
CART



Accuracy: 0.875
SENSITIVITY (aka RECALL): 0.766
SPECIFICITY (aka FALL-OUT): 0.915
PRECISION (aka POSITIVE PREDICTIVE VALUE): 0.766
NEGATIVE PREDICTIVE VALUE): 0.915



Logistic Regression



Accuracy: 0.983
SENSITIVITY (aka RECALL): 0.979
SPECIFICITY (aka FALL-OUT): 0.984
PRECISION (aka POSITIVE PREDICTIVE VALUE): 0.958
NEGATIVE PREDICTIVE VALUE): 0.992

TEST ON TEAM MEMBERS

Team Member	Traits of Autism
1	N
2	N
3	N
4	Y
5	N

SUMMARY

- (Q9) The inability to understand what someone is thinking, just by looking at their face is a variable which is most significantly related to traits of ASD.
- We cannot recommend someone for a full ASD assessment on the basis of their Age.
- Jaundice, Relatives with Autism, and ethnicity
- LOGIT performed better than CART

RECOMMENDATIONS

- Previously the threshold is 0.6. But we do not want to leave anyone that might be autistic out. Based on our model, the new threshold is 0.343. we can calculate the probability of someone is Autistic, apply the threshold then determining this person is on the spectrum or not. We are using more data to predict, his way minimizing false negative. More people can be diagnosed and treated
- We can use some of the data to create marketing messages, raise awareness, from people who are in this situation to take the ASD test.
- We recommend policy maker, insurance company to get more, related data and take action such as providing them access to assessment.

