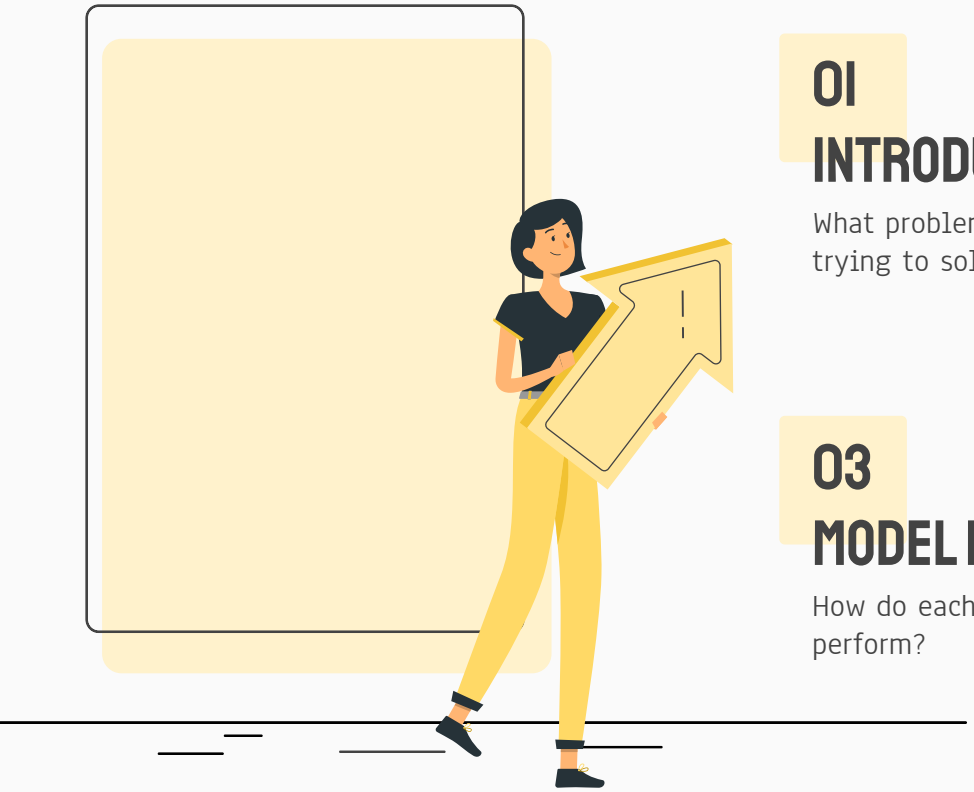
CLASSIFYING AUTISM

Ethan van Heerden, Karen Phung, Kevin Lee



01 INTRODUCTION

What problem are we trying to solve?

02 METHODOLOGY

What classification methods can we use?

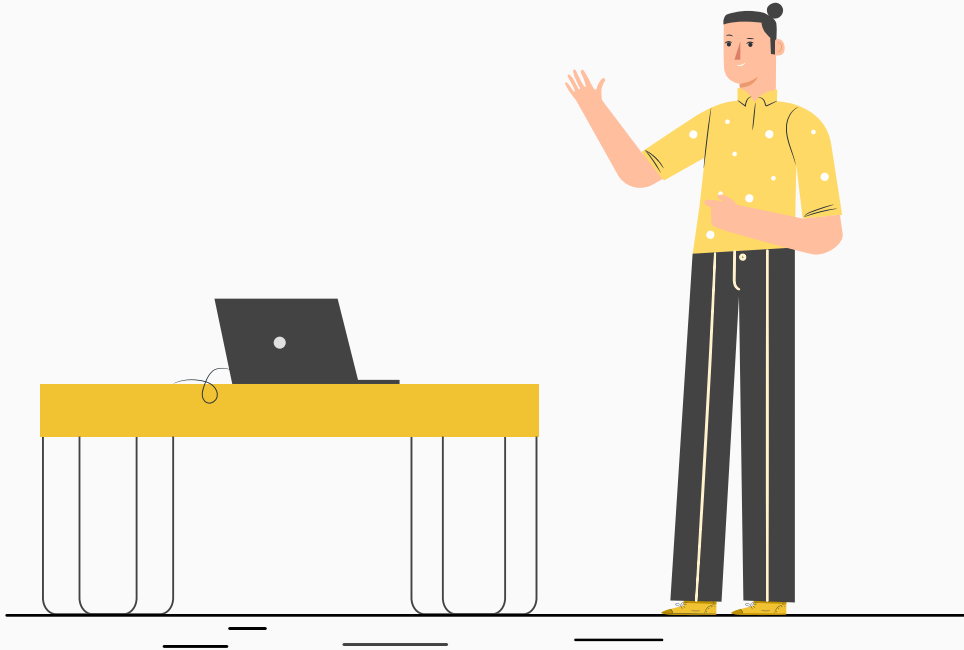
03 MODEL EVALUATION

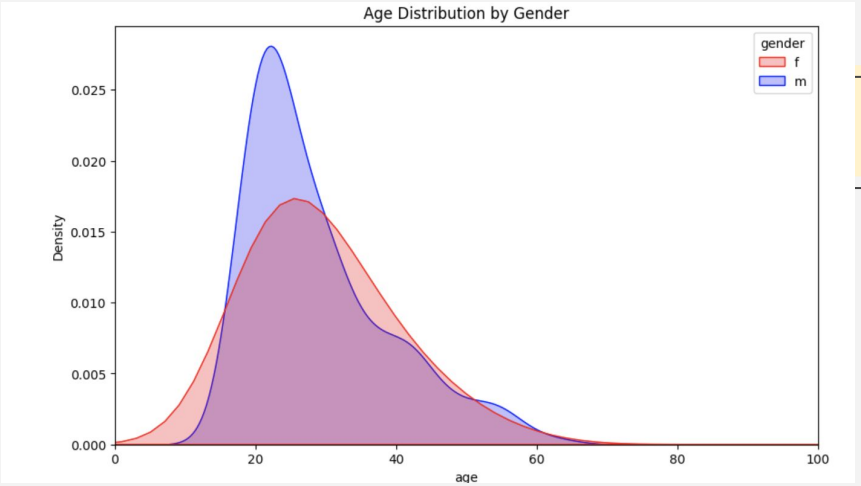
How do each of our models perform?

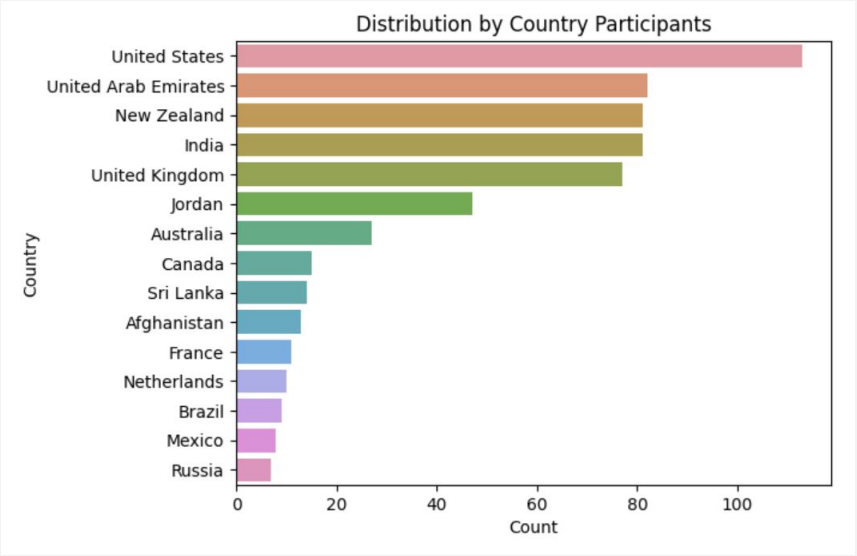
04 CONCLUSIONS

Which model was best?

01.
INTRODUCTION





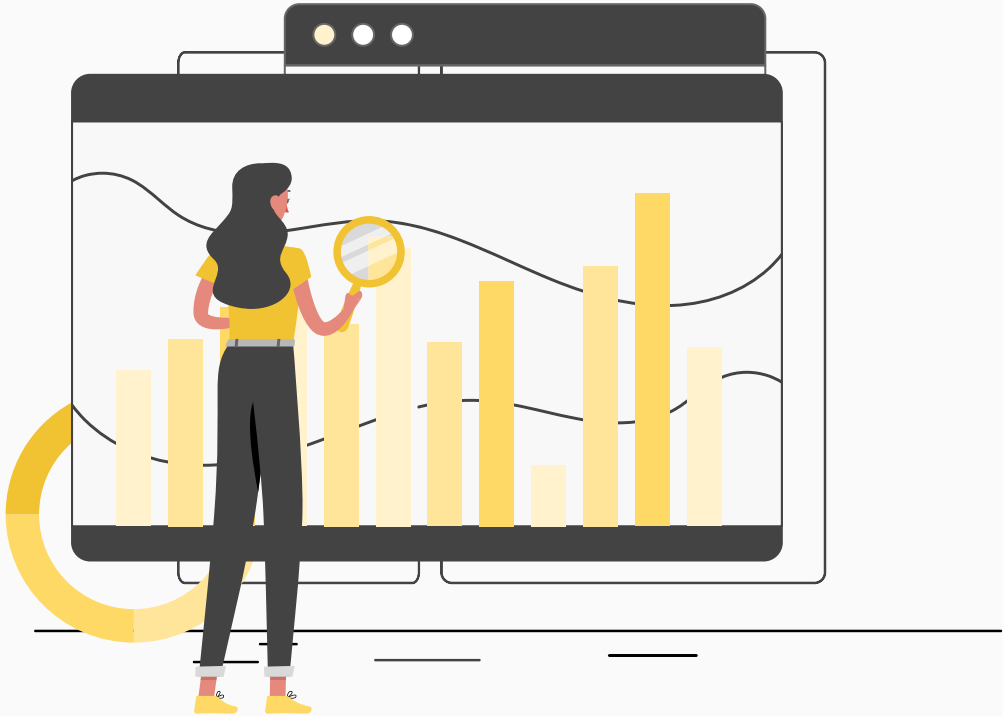


INTRODUCTION & DATASET

Feature	Type	Description
Age	Number	Toddlers (months), children, adolescent, and adults (year)
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician,etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening method type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0 = toddler, 1 = child, 2 = adolescent, 3 = adult)
Language	String	(English, Arabic, Farsi, Mandarin, Urdu, Swahili, French, Spanish, Portuguese, Turkish)
Why_are_you_taken_the_screening	String	Use input textbox
Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner
Class	String	ASD traits or No ASD traits (automatically assigned by the ASDTests app).

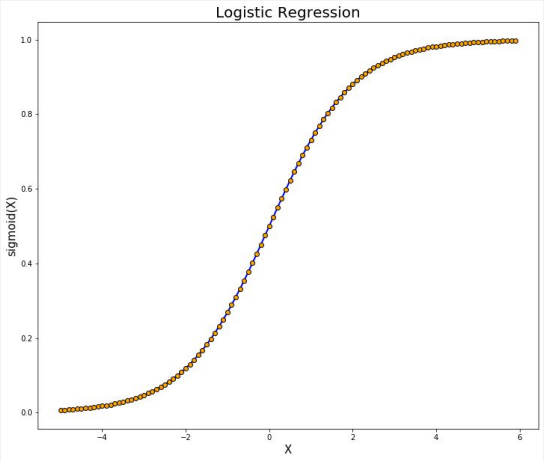
Thabtah F, Abdelhamid N, Peebles D. A machine learning autism classification based on logistic regression analysis. Health Inf Sci Syst. 2019 Jun 1;7(1):12. doi: 10.1007/s13755-019-0073-5. PMID: 31168365; PMCID: PMC6545188.

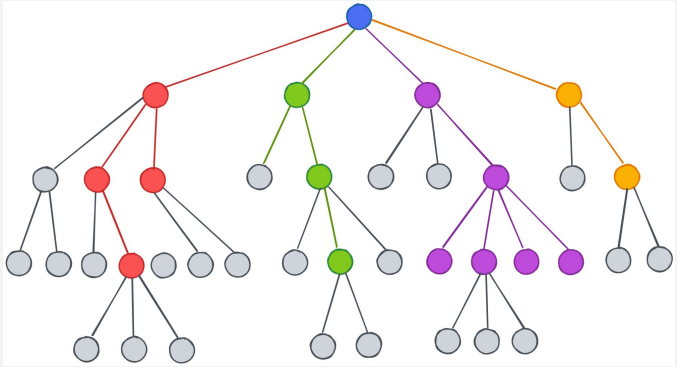
02.
METHODOLOGY

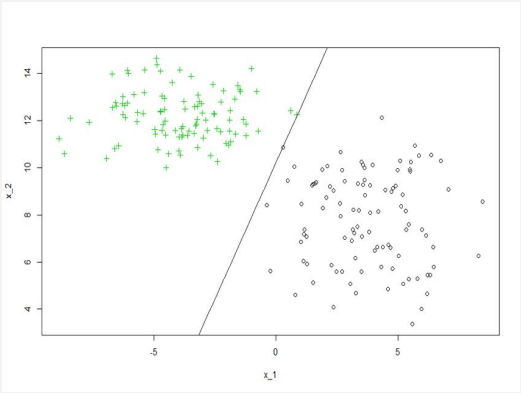


MODELING TECHNIQUES

- In terms of bias, the three models show varying degrees. Logistic Regression has the highest bias, followed by the Perceptron, and the Decision Tree has the lowest bias.
- Our choice of models were based on comparing their classification performance. We decided to choose Logistic Regression, Perceptron, and Decision Tree due to them being popular classification models.







DATA COMPLEXITY AND PREPROCESSING

The researcher's data contains a mixture of numerical, binary, and categorical values

NUMERICAL

- 12 features
- **SimpleImputer** to deal with missing values
- **StandardScaler** to scale everything

BINARY

- 2 features
- Custom transformer to convert "no" to 0 and "yes" to 1

CATEGORICAL

- 5 features
- **OneHotEncoder** to transform each category into a one-hot numeric array

03.
MODEL EVALUATION

DECISION TREE

Tuned hyperparameters using **recall** as our scorer

01
MAX DEPTH

- Controls the maximum allowed depth of the tree
- Best value: 1

02
MIN SAMPLES FOR SPLIT

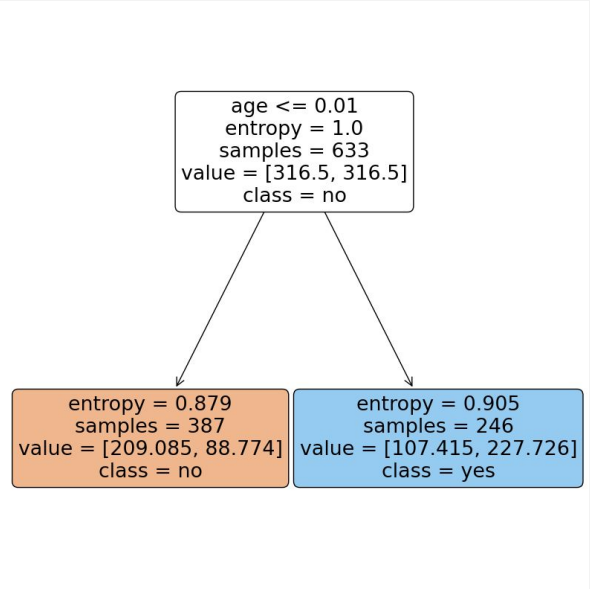
- Controls the minimum number of samples required to split an internal node
- Best value: 2

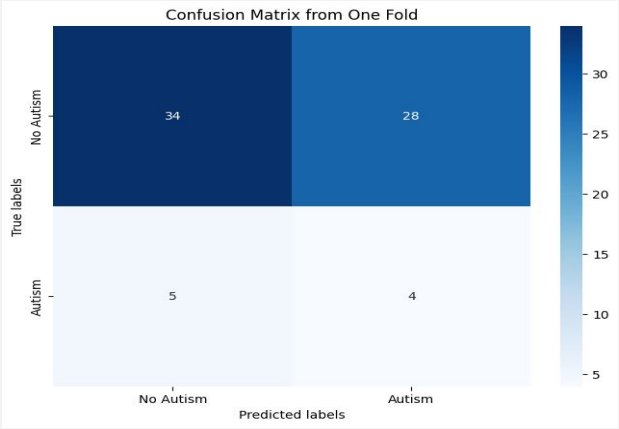
03
MAX CONSIDERED FEATURES

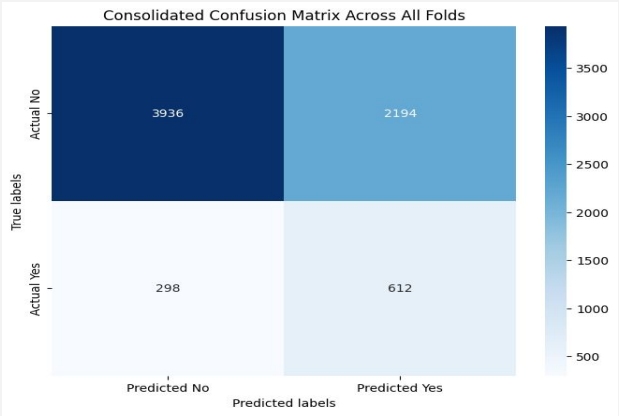
- Controls the number of features to consider when looking for the best split
- Best value: 0

DECISION TREE RESULTS

ACCURACY	0.646
⭐RECALL	**0.669**
PRECISION	0.218
FI-SCORE	0.326







PERCEPTRON

Tuned hyperparameters using **recall** as our scorer

01
ALPHA

- The regularization term
- Best value: 0.0001

02
MAX_ITER

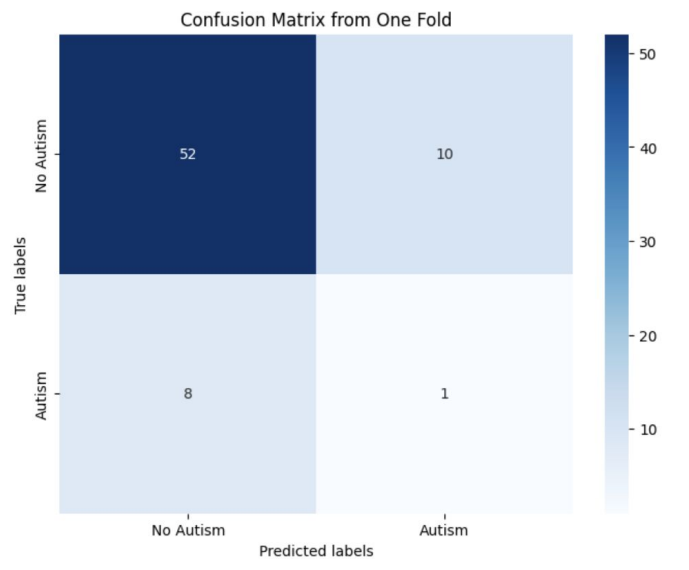
- The maximum number of passes over the training data.
- Best value: 500

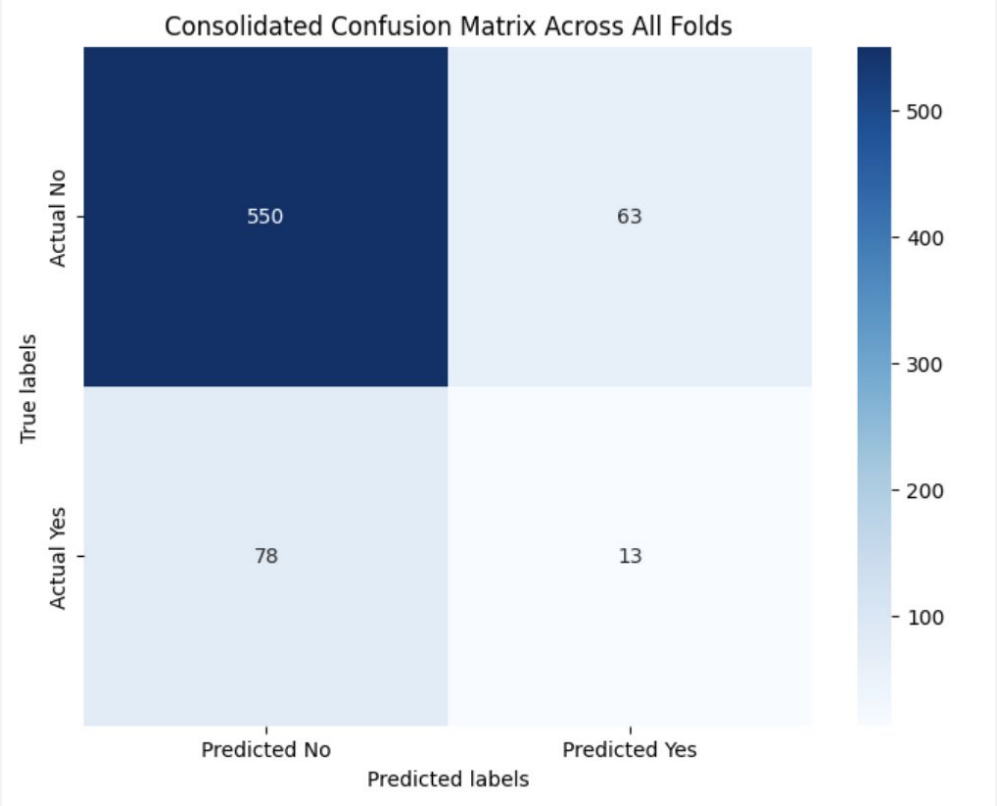
03
ETA0

- The initial learning rate
- Best value: 0.001

PERCEPTRON RESULTS

ACCURACY	0.799
⭐RECALL	0.139
PRECISION	0.309
FI-SCORE	0.131





LOGISTIC REGRESSION

Tuned hyperparameters using **recall** as our scorer

01
C-VALUE

- The regularization strength
- Best value: 0.0336

02
MAX_ITER

- The maximum number of passes over the training data.
- Best value: 100

03
PENALTY

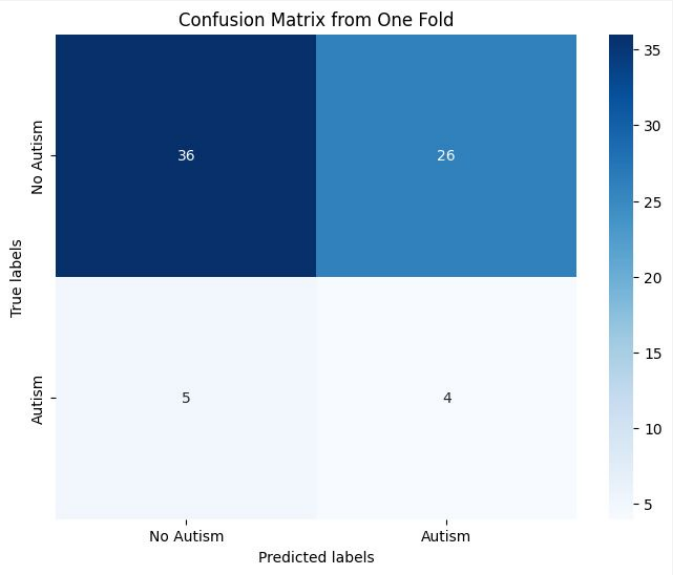
- The regularization type
- Best type: L1

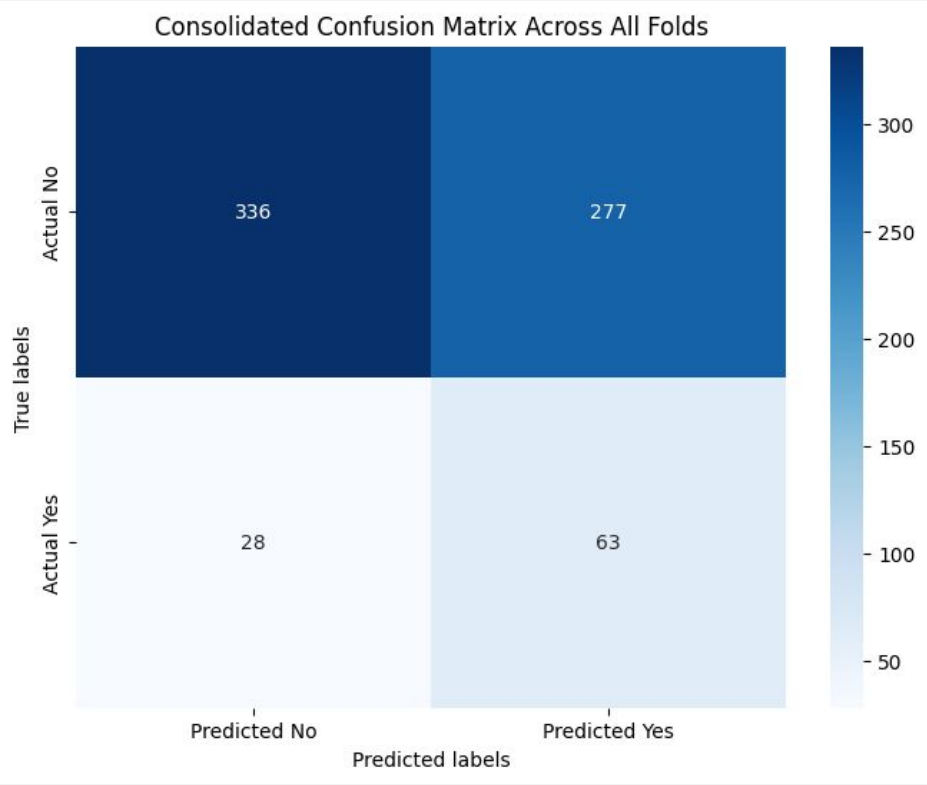
04
SOLVER

- The optimization algorithm
- Best algorithm: liblinear

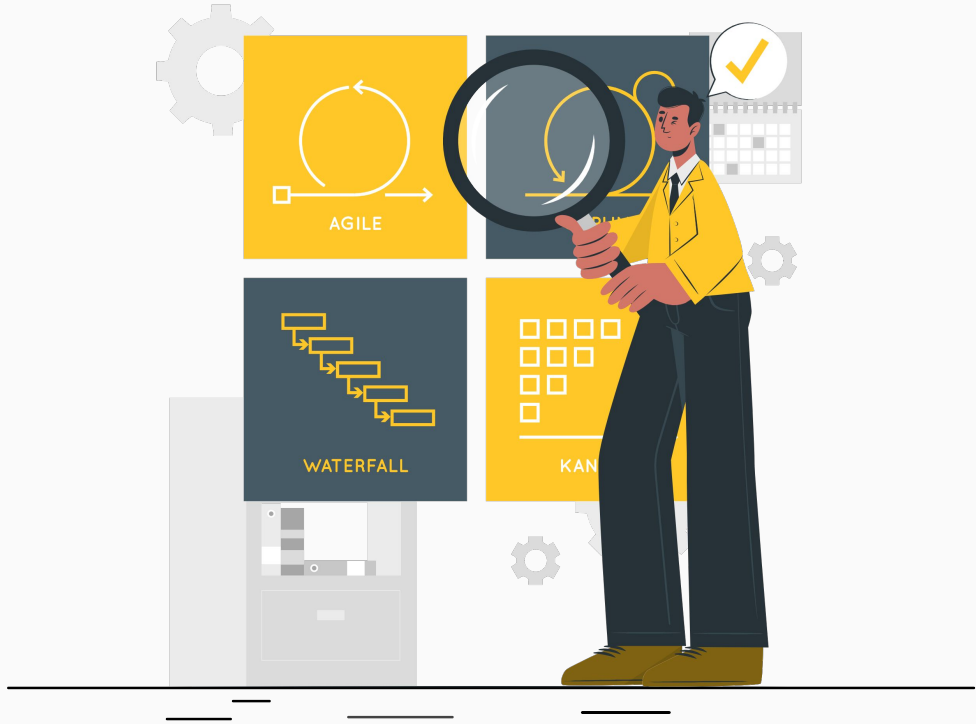
LOGISTIC REGRESSION RESULTS

ACCURACY	0.567
⭐RECALL	0.690
PRECISION	0.184
FI-SCORE	0.290





04.
CONCLUSIONS



CONCLUSION

COMPARING THE DIFFERENT MODELS

🥇 **Logistic Regression** - Recall of 0.690

🥈 **Decision Tree** - Recall of 0.669

🥉 **Perceptron** - Recall of 0.139

POSSIBLE IMPROVEMENTS WITH EXISTING DATA

- Use SVMs and the kernel trick

FUTURE WORK

- Add more data to fix imbalances (613 no vs. 91 yes)
- More features?
- Research the effects of age on autism
- Obtain data from children who have autism

THANK YOU

Any questions?