Prediction of Autism

Fatemeh Kamiab

Prediction of Autism is valuable for Healthcare

- Due to the complexity, variety and the closeness of the symptoms of Autism and other developmental disorders, the diagnosis of Autism is not easily possible.
- Waiting times for an ASD diagnosis are lengthy and procedures are not cost effective.
- The economic impact of Autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods.
- Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis.

Data acquisition and cleaning

- **Data Type**: Predictive and Descriptive: Nominal / categorical, binary and continuous
- **Task**: Classification
- **Attribute Type:** Categorical, continuous and binary
- **Area:** Medical, health and social science
- **Format Type:** Non-Matrix
- **Does your data set contain missing values?** No
- **Number of Instances (records in your data set):** 1054
- **Number of Attributes (fields within each record):** 18 including the class variable
- **Attribute Information:** For Further information about the attributes/feature see below table

Features collected and their descriptions

Feature	Type	Description
A1: Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A2: Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A3: Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A4: Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A5: Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A6: A6: Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A7: Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A8: Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A9: Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
A:10 Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Age	Number	Toddlers (months)
Score by Q-chat-10	Number	1-10 (Less that or equal 3 no ASD traits; > 3 ASD traits
Sex	Character	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 ASD history	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.
Why_are_you_taken_the_screening	String	Use input textbox
Class variable	String	ASD traits or No ASD traits (automatically assigned by the ASDTests app). (Yes / No)

Solution to the problem

The goal was to find a “best-fitting” model using logistic regression . The ability to predict the probability (risk) of a child having ASD through some traits could greatly contribute to the difficult diagnosis process currently in place.

Modeling

Let’s build our model using **Logistic Regression** from Scikit-learn package

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
LR = LogisticRegression(C=0.01, solver='liblinear').fit(X_train,y_train)
LR
```

```
LogisticRegression(C=0.01, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=None, solver='liblinear', tol=0.0001, verbose=0,
                   warm_start=False)
```

Performances of different Evaluation

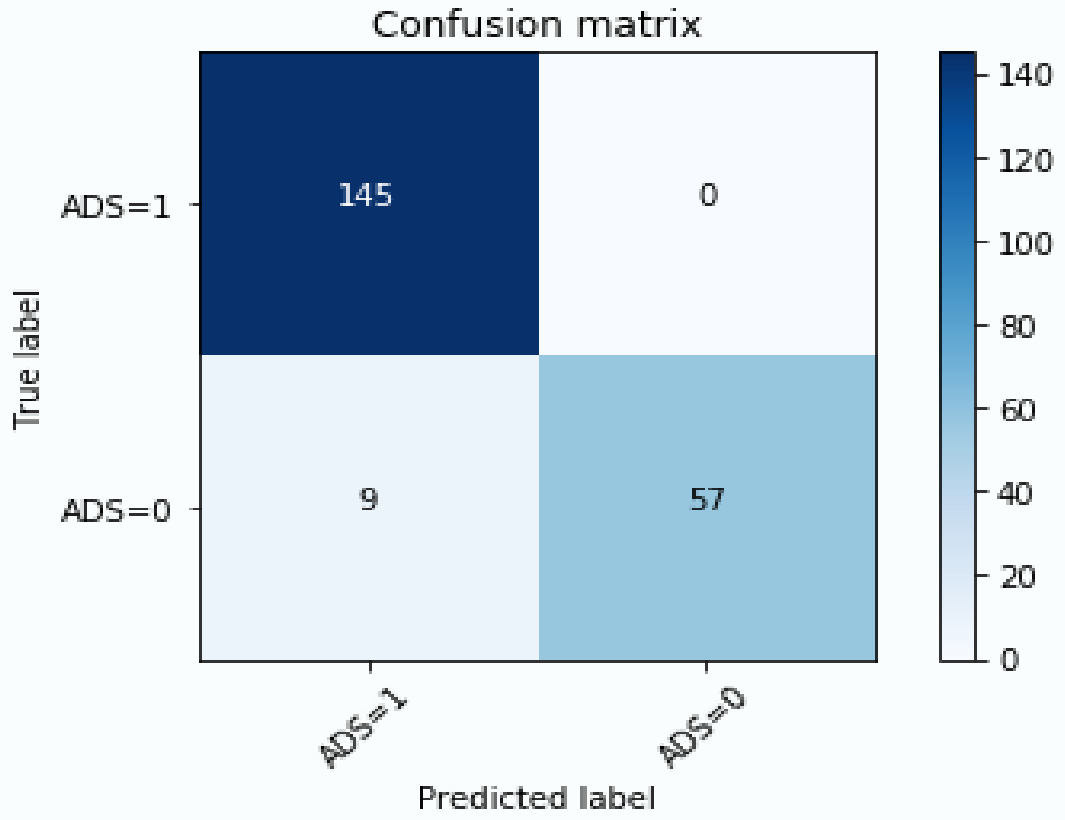
jaccard index

```
from sklearn.metrics import jaccard_score
jaccard_score(y_test, yhat)
```

```
0.9415584415584416
```

Performances of different Evaluation

Confusion matrix



Performances of different Evaluation

Classification's Report

```
print (classification_report(y_test, yhat))
```

	precision	recall	f1-score	support
0	1.00	0.86	0.93	66
1	0.94	1.00	0.97	145
accuracy			0.96	211
macro avg	0.97	0.93	0.95	211
weighted avg	0.96	0.96	0.96	211

Performances of different Evaluation

log loss

```
from sklearn.metrics import log_loss
log_loss(y_test, yhat_prob)
```

```
0.25636419958456347
```

Conclusion

- This study was especially limited due to sample size, and even a large sample would still require a screening of the extensive list of predictor variables to identify potentially “important “predictors.
- Study of the various elements in the sample’s behavior with a clinical professional is important in terms of finding an adequate starting point with which to begin model selection.
- The ability to obtain a larger sample would allow for the consideration of more predictors into the multivariate model.

Future work

- Though the model obtained worked well for the current sample, there is a possibility that the variation observed within this sample is not a good representation of the population as a whole.
- A first step toward future work would be to validate this model repeatedly for other sample data to see how well it performs.
- Second, and possibly most important, would be to increase the sample size if possible so the model is not as sensitive to influential data and over-fitting. Outliers and influential data had much more of an impact in this study in regards to decision making to preserve sample size.
- Lastly, an in depth discussion with a clinical professional about the clinical importance of the variables both in the multivariate model and even those excluded from the model could have a significant impact in the model building process.