CAPSTONE PROJECT – STATISTICAL DATA MINING FOR BIG DATA

Using Machine Learning Methods to Detect Autism Spectrum Disorder in Adults

William Belcher

Autism Spectrum Disorder is a neurodevelopment condition in which the people with the condition exhibit a variety of distinct behavioural patterns. Diagnosis typically happens at a young age as parents and teachers notice the autism symptoms. However, identifying autism in adults is particularly difficult as its symptoms overlap with a variety of other mental health conditions. With machine learning techniques growing ever popular in the clinical setting, a rapid screening process that could be used in assisting the referral of patients to medical professionals was essential. In this paper, the possibility of using Logistic Regression, Naïve Bayes and tree-based classifiers is explored. The proposed methods were evaluated on a publicly available AQ-10-Adult based screening data set that contained 704 observations of 21 attributes. After the relevant pre-processing of the data and application of the above-mentioned techniques, the achieved results indicated that a binary logistic regression model on a reduced data set containing the first two principle components derived from the AQ-10 questionnaire and a subset of the environmental predictors provided the best cross-validated accuracy of 98.56%, a test accuracy of 99.19% and a sensitivity of 100%. From the results achieved, it can be argued that the use of a binary logistic regression-based model in the clinical setting can assist in identify autism spectrum disorder in adult patients.

Introduction

Autism Spectrum Disorder (ASD) is a neurological condition associated with many atypical mannerisms and behavioural patterns, most notably those surrounding interpersonal interactions. Autism Spectrum Disorder is a condition related to the development of the human brain. It is worth noting that both environmental and genetics may be contributing factors in the development of ASD. However, scientists have been unable to uncover the root cause and as such ASD is usually detected through observations and diagnosed by a specialist. Unfortunately, the process for receiving an ASD diagnosis are lengthy with multiple appointments with specialists not being cost effective. Early detection of the condition can assist in the improvement of the subject's overall health by enabling them to implement techniques and medication that reduces the impact of the condition on their daily lives sooner. With the rapid increase in modern computing power and number of machine learning models assisting in the diagnosis of medical conditions, the early detection of ASD based on variety of physiological attributes now seems viable. The detection of autism spectrum disorder in a patient proves difficult as, as the name implies, the disorder is a spectrum resulting in significant intragroup variance in those being classified as having the condition. A time-efficient and easily accessible screening process is necessary in assisting medical professionals in informing individuals whether they should pursue a formal, clinical diagnosis.

Data

The Autism Screening Data for Adults data set, collected from the UCI Machine Learning Repository, contains several predictors and one target variable (Thabtah, 2021). The data was originally collated by the Manukau Institute of Technology for use in the development of a time efficient and accessible ASD screening process. The response variable classified observations into two categories: ASD and Non-ASD. The first ten attributes were the person's responses to the AQ-10-Adult Questionnaire (Allison, Auyeung & Baron-Cohen, 2012). The remaining factors were related to the patient's childhood and current environment. These variables contained information on the patients age, gender, ethnicity, whether they suffered from jaundice as a child, their current country of residence, whether they had used the screen app before, and their age group. The dataset's original twenty attributes are listed in Table 1:

Attribute ID	Description
1-10	Answer to the corresponding AQ-10-Adult questionnaire
11	Patient age (years)
12	Gender (m/f)
13	Ethnicity
14	Did the patient have jaundice at birth? (yes / no)

15	Family history of ASD (yes / no)
16	Patient country of residence
17	Has the patient used the screening app before? (yes / no)
18	Screening score
19	Age group (18 or older)
20	Person who's using the screening app's relation to patient

Table 1: Attribute list

In total, there were 704 observations.

Data Pre-processing

Data pre-processing is a technique in which the raw data is transformed into a meaningful and understandable format. The 'Age group' column was removed from the data as it only contained one value of '18 and over'. The 'Screening score' column was also removed as it was just the sum of the answers to the AQ-10 questions. The "Used screening app before" as well as the "relation to patient" columns were removed as they are deemed unnecessary by inspection. The "country of residence" was also not included in the model training data as it was deemed unimportant for the scope of this task. Missing values in the data are denoted with a '?' with rows containing missing values being removed from the data. Missing value imputation was not conducted in order to not introduce unnecessary variance in the predictors. A single outlier was detected and removed. The "patient ethnicity" column contained 11 categories. In order to reduce the dimensionality of the data, these categories were collapsed in 4, more general, categories being: "White", "Asian", "Black", "Other". These new categories were similarly sized with proportions of 38%, 26%, 23% and 13% of the total number of observations respectively. Figure 1 illustrates these proportions of the new categories. The size of the cleaned data used for model generation contained 17 columns, 16 predictors and 1 response, with 608 total observations

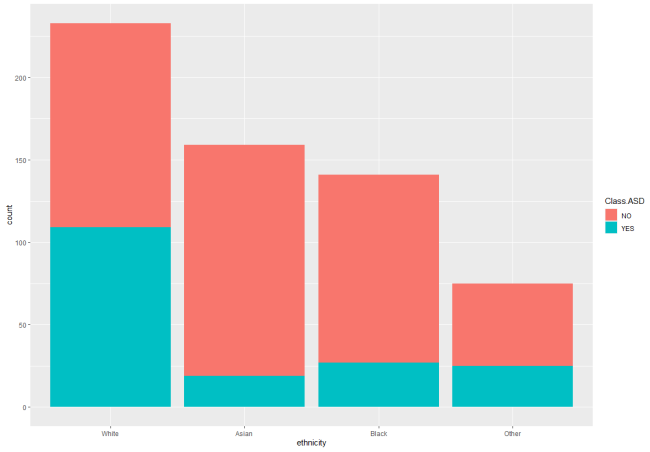

Figure 1: Collapsed Ethnicity Proportions

Methods

Principle Component Analysis (PCA)

Principal component analysis is an unsupervised learning technique used for dimension reduction as well as exploratory data analysis by projecting the variables onto a new, orthogonal basis that can be used to illustrate the proportion of variance explained by each principal component. The k-th principle component can be found by subtracting the first $k-1$ components from $\boldsymbol{X}$:

$$\widehat{\boldsymbol{X}_k} = \boldsymbol{X} - \sum_{i=1}^{k-1} \boldsymbol{X}\boldsymbol{w}_i\boldsymbol{w}_i^T$$

And then finding the vector of weights which extracts the maximum variance from the resulting matrix:

$$\boldsymbol{w}_k = \max_{\|\boldsymbol{w}\|=1} \left\{ \left\| \widehat{\boldsymbol{X}_k}\boldsymbol{w} \right\|^2 \right\} = \max_{\|\boldsymbol{w}\|=1} \left\{ \frac{\boldsymbol{w}^T \widehat{\boldsymbol{X}_k}^T \widehat{\boldsymbol{X}_k}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}} \right\}$$

Training and testing split

The complete data set was split into training and testing subsets using an 80/20 split. K-fold validation with the training subset and with k = 5 was used for each model. Figure 2 shows the final training, validation and testing sets that were used.

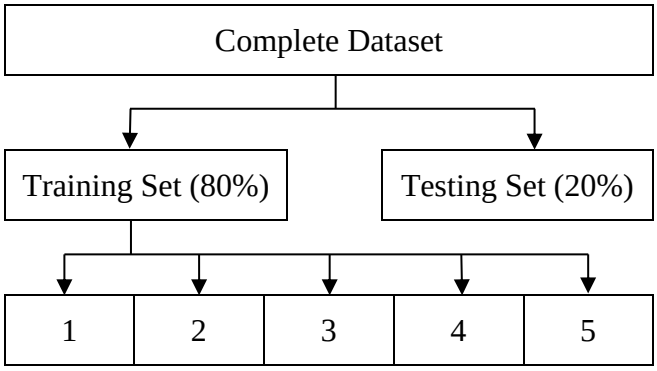

Figure 2: Data split

Logistic Regression (LR)

Logistic regression is based off the standard linear regression methods but applies the logit transformation to the $p(x)$ resulting in a response that will lie between the value of 0 and 1 inclusive.

$$logit(\pi_i) = log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Naïve Bayes (NB)

The naïve Bayes classifier is a probabilistic machine learning model that is based on Bayes theorem, finding the probability of y happening given that X has already occurred. The naivety of the model comes from the assumption that features are independent in each class.

$$P(y|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|y)P(y)}{P(\boldsymbol{X})}$$

Classification Tree (CT)

Classification trees are a subtype of decision trees where the target variable can only take a discrete set of values. In the tree structures, the leaves represent the class labels with the branches representing conjunctions of features that lead to those class labels.

Random Forest (RF)

The random forest algorithm is an extension of the classification tree that fits many classification trees to a data set and the combines the predictions from all of the trees (Mubayi, 2017). The prediction from the algorithm can be explained with the following equation:

$$\hat{C}_{rf}^{B}(x) = majority\ vote\left\{\hat{C}_b(x)\right\}_1^B$$

Where $\hat{C}_b(x)$ is the prediction from the b-th tree in the random forest.

Best Subset Selection

Best subset selection is a method that aims to find the subset of independent predictors that best predict the response and does so by comparing all possible combinations of the predictors. This method works well for data with small dimensions, but as the predictors increase linearly, the possible number of combinations increases exponentially. The model selection criteria utilised in best subset is the Akaike's Information Criteria (AIC). AIC attempts to find the distance between the true likelihood function of the given data and the likelihood function of the fitted model plus a constant to penalise over complex models.

$$AIC = -2\,\ell\left(\hat{\beta}_M\right) + 2p$$

Where p is the number of parameters in model M.

Receiver Operating Characteristic (ROC)

The receiver operating characteristic curve is a graphical plot the illustrates the diagnostic ability of a binary classification system. The curve is created by plotting the true positive rate against false positive rate for varying thresholding values. Additionally, the area under the curve (AUC) provides an aggregate measure of performance across all the threshold values ("Classification: ROC Curve and AUC | Machine Learning Crash Course", 2021). A high sensitivity indicates that the model was able to successfully identify those patients with autism as having autism. Subsequently, a high specificity indicates that the model was able to correctly classify those without ASD has not having ASD. Generally speaking, a model with a sensitivity and specificity around 90% are considered to have good diagnostic performance (Hoffman, 2021).

Software Packages

The pre-processing and machine learning techniques were completed using the software package RStudio, an integrated development environment for the R programming language (RStudio: Integrated Development Environment for R, 2021).

Performance Metrics

The result of a model is measured in terms of its specificity, accuracy and sensitivity. The values are obtained by using the generated models to predict the outcome on the test data set and making note of the resulting confusion matrix as seen in Table 2 and Table 3:

	Actual Outcome	
Predicted Outcome	True Positive	False Positive
	False Negative	True Negative

Table 2: Confusion Matrix

Performance Metrics
$Accuracy = \frac{TP + TN}{TN + TP + FN + FP}$
$Specificity = \frac{TN}{TN + FP}$
$Sensitivity = \frac{TP}{TP + FN}$

Table 3: Performance Metrics formulae

Results and Discussion

Principle component analysis was conducted on the AQ-10-Adult questions, in order to find the minimal number of questions required to explain a significant amount of the variance in the data. PCA illustrated that there was no significant 'elbow' in the proportion of variance explained by each principal component. However, with only 3 components, approximately 50% of the variance in the response could be explained as seen in Figure 3.

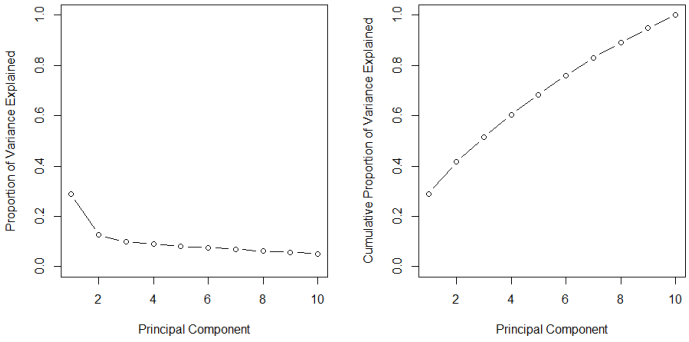

Figure 3: Proportion of Variance Explained, Scree Plot

These three components, along with the patient's age, gender, collapsed ethnicity, family history of autism and whether they suffered from jaundice as a child were used in training the following models.

Best subset selection was first completed on the logistic regression model. This returned a model which only contained the three principle components, the age and whether the patient suffered from jaundice as a child. The predictors found to be significant in each subset can be found in Appendix A. Of these five predictors, only PC1 and PC3 were found to be statistically significant.

Creating an LR model with the response being predicted purely by PC1 through to PC3 achieved an accuracy of 99.18% on the test set. This high accuracy from only two principle components can be attributed to the way ASD is diagnosed. Figure 4 illustrates the data projected onto the first two principle components and reveals distinct clusters exposed by PCA.

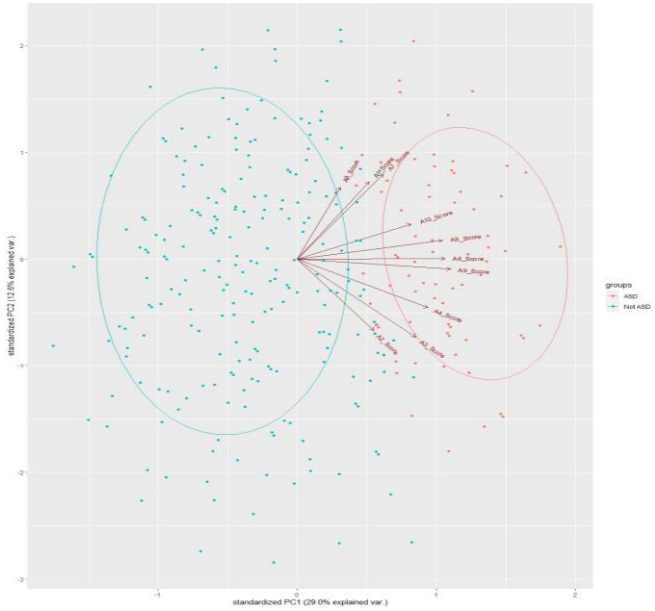

Figure 4: Graphical plot of PC1 and PC2

A psychiatrist will make their diagnosis after looking exclusively at the behavioural patterns displayed by the patient which the AQ-10-Adult questionnaire aims to identify. The psychiatrist does not consider other aspects of the patient such as their age, gender, etc.

The logistic regression and naïve bayes models were trained with the predictors found to be significant from the best subset selection. The tree-based classifiers were provided the full training data as the R functions used only utilised the features it found to be imperative. Figure 5 illustrates the range of cross validated errors produced from each model with:

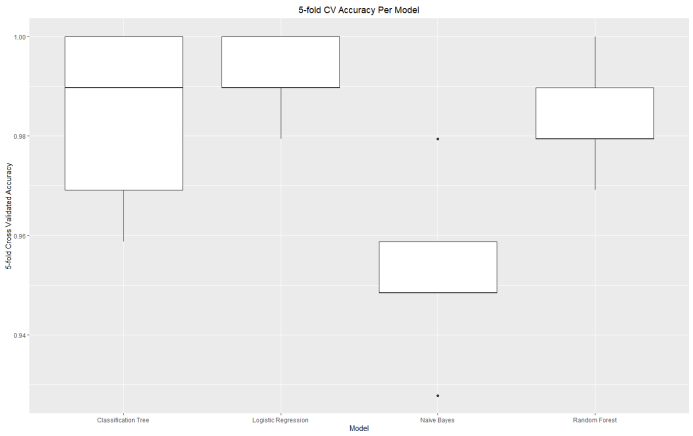

Figure 5: Accuracy from 5-Fold CV per Model

From the figure, it can be seen that the classification tree, logistic regression and random forest models proved to be both accurate and stable with the logistic regression model being the most consistent. Figure 6 illustrates the resulting classification tree:

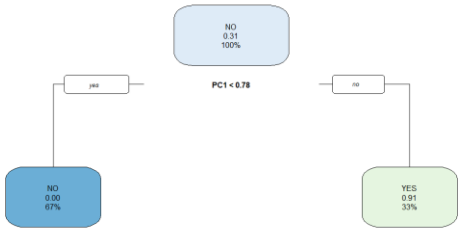

Figure 6: Generated Classification Tree

As illustrated, the classification tree only utilised PC1. Again, this is related to the way in which ASD is diagnosed. All models were then tested with the 20% test set. Table 4 provides the full results from all the models tested:

Model	Accuracy	Sensitivity	Specificity
LR	99.19%	100%	98.81%
NB	98.37%	100%	97.65%
CT	95.59%	86.11%	100%
RF	98.14 %	96.05%	99.10%

Table 4: Testing Results per Model

The result from the testing set states that logistic regression was most accurate at 99.19%. The 100% sensitivity for both the logistic regression and the naïve bayes models is imperative, as misclassifying someone with ASD as not having ASD will lead to significant physiological and economical detriments for the patient. The receiver operating characteristic for each model can then be compared. This plot visualises the relationship between sensitivity and specificity for a given model. Figure 7 illustrates these relationships:

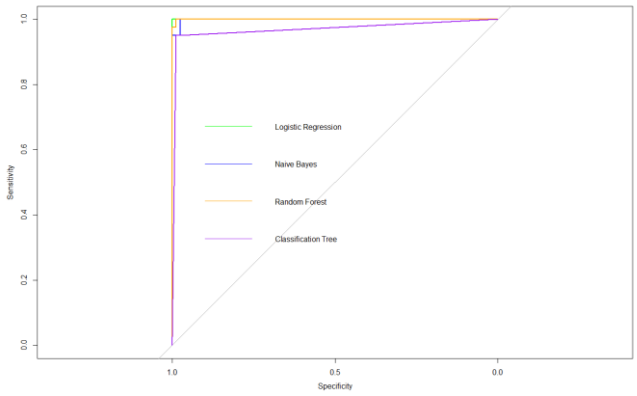

Figure 7:ROC Curve per Model

As seen in the figure, all the models had an extremely strong relationship between the sensitivity and specificity, with all of them being extremely close to 100% in both categories.

Conclusion

To conclude, the use of machine learning models in a clinical setting can assist in the early detection of autism spectrum disorder in adults. Principle component analysis revealed that by utilising only 3 principle components derived from the AQ-10 questionnaire a high accuracy and 100% sensitivity can be achieved. This project proposes the use of a logistic regression-based model that can assist in the diagnosis of ASD through the use of the AQ-10-Adult questionnaire and a handful of other factors such as the patient age and whether they suffered from jaundice as a child. The proposed model had a cross-validated test accuracy of 98.56% and a test accuracy of 99.19%. This model also had a sensitivity of 100% and a specificity of 98.81%. These results strongly suggest that a binary logistic regression-based model will be able to accurately predict autism spectrum disorder in adults, reducing the time taken and therefore reducing the financial impact of the diagnosis process. This reduction in diagnosis time enables people with the condition to implement techniques and medication that reduces the impact of the condition on their daily lives sooner.

References

- Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward Brief "Red Flags" for Autism Screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 Cases and 3,000 Controls. Journal Of The American Academy Of Child & Adolescent Psychiatry, 51(2), 202-212.e7. doi: 10.1016/j.jaac.2011.11.003

- Classification: ROC Curve and AUC | Machine Learning Crash Course. (2021). Retrieved 18 October 2021, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

- Hoffman, R. (2021). Understanding medical tests: sensitivity, specificity, and positive predictive value. Retrieved 23 October 2021, from https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/understanding-medical-tests-sensitivity-specificity-and-positive-predictive-value

- Mubayi, A. (2017). Computational Modeling Approaches Linking Health and Social Sciences : Sensitivity of Social Determinants on the Patterns of Health Risk Behaviors and Diseases. Handbook Of Statistics, 36, 249-304. doi: 10.1016/bs.host.2017.08.003

- PBC. (2021). RStudio: Integrated Development Environment for R (Version 1.4.1717) [Windows]. Boston, MA.

- Thabtah, F. (2021). UCI Machine Learning Repository: Autism Screening Adult Data Set. Retrieved 1 October 2021, from https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult

Appendix A - Significant Predictors per Subset

# of parameters / predictors	0	1	2	3	4	5*	6	7	8
Intercept	■	■	■	■	■	■	■	■	■
PC1		■	■	■	■	■	■	■	■
PC2			■	■	■	■	■	■	■
PC3				■	■	■	■	■	■
Age						■	■	■	■
Gender							■	■	■
Ethnicity					■			■	■
Jaundice						■	■	■	■
Autism									■
Loglikelihood	-2.9e+2	-3.3e+1	-2.3e+1	-1.0e+1	-2.3e-4	-2.4e-5	-1.9e-5	-4.5e-7	-4.5e-7
AIC	598	68	50	26	12	10	12	18	20